Univerzita Karlova v Praze, Filozofická fakulta

Ústav českého jazyka a teorie komunikace

Filologie - Český jazyk

Kvantitativní charakteristiky termínů

Dizertační práce

Quantitative Characteristics of Terms

Dominika Kováříková

školitel: Prof. PhDr. František Čermák, DrSc.

2014

Prohlašuji, že jsem dizertační práci napsala samostatně s využitím pouze uvedených a řádně citovaných pramenů a literatury a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

Mgr. Dominika Kováříková

Abstrakt

Metoda automatického vyhledávání termínů TERMIT je zaměřena nejen na samotnou úspěšnost, tedy co nejvyšší počet správně vyhledaných termínů, ale v první řadě na vlastnosti, které při identifikaci jednoslovných a víceslovných termínů hrají nejdůležitější roli. Je založena na data miningu, tedy na vytěžování informací z velkých objemů (korpusových) dat. Metoda TERMIT se při rozpoznávání termínů v reálných textech i při hledání podstatných kvantitativních rysů termínů osvědčila. Na jejím základě je možné jednoslovný termín charakterizovat jako slovo, které se v odborných textech daného oboru vyskytuje výrazně častěji než v textech neakademických, vyskytuje se jen v malém počtu akademických disciplín, v celém korpusu (SYN2010) je nerovnoměrně rozložené a málo frekventované a rozestupy mezi jeho jednotlivými výskyty jsou nepravidelné. Víceslovný termín je podle výsledků metody TERMIT ustálená kolokace složená z méně frekventovaných slov, která obvykle obsahuje alespoň jedno slovo s vysokou terminologickou platností, tedy jednoslovný termín. S pomocí těchto charakteristik termínů lze více než 95 % textu zařadit správně mezi jednoslovné i víceslovné termíny a netermíny. Na základě automaticky vyhledaných termínů lze potom například zjišťovat množství jednoslovných i víceslovných termínů v textech různých disciplín či sledovat vztahy mezi obory na základě sdílených termínů. Obecně lze konstatovat, že automatické vyhledávání jazykových jevů může přispět k jejich bližšímu poznání a že i čistě kvantitativní přístup, jako je data mining, je vhodný pro zkoumání lingvistického (korpusového) materiálu.

Abstract

The new method of automatic term recognition TERMIT is focused not only on the high number of correctly labeled terms, but also on the most important attributes of a term (in terms of their role in automatic term identification process). The method is based on data mining, i.e. finding meaningful information in very large corpus data. It was able to both successfuly identify terms in academic texts and find constitutive features of a term as a terminological unit. The single-word term (SWT) can be characterized as a word with a low frequency in corpus (SYN2010) that occurs considerably more often in specialized texts of a given field than in non-academic texts, occurs in a small number of academic disciplines, its distribution in the corpus (SYN2010) is uneven as is the distance between its two instances. The multi-word term (MWT) is a stable collocation consisting of words with low frequency and contains at least one (and often more) single-word term. Based on the characteristics of SWT and MWT, it is possible to classify individual tokens in texts as terms or non-terms with a success rate of more than 95 %. Automatically identified terms can be used to identify percentage of SWT or MWT in different academic disciplines, as well as find terms shared by two or more domains in order to assess their relationship. In general, we can conclude that an automatic recognition of a languge phenomenon can contribute to its characterization and that a purely quantitative approach (such as data mining) can be used to research linguistic (corpus) material.

Poděkování

Poděkování patří především prof. F. Čermákovi za inspiraci a za obětavé vedení dizertační práce. Zvláštní díky náleží ing. O. Kováříkovi za pomoc a radu při technickém zpracování a za neustálou podporu. Děkuji i doc. V. Cvrčkovi za nesčetné konzultace jednotlivých problémů.

Za odbornou terminologickou radu děkuji všem konzultantům z nelingvistických oborů: ing. O. Kováříkovi (obor informatika), Mgr. O. Hanusovi (obor literární věda), Mgr. P. Pabinovi, PhD (obor sociologie) a MUDr. R. Čáberovi (obor medicína).

Obsah

Úvod

Cílem předkládané dizertační práce je získat nové poznatky o termínech, zvláště o jejich vlastnostech a chování v textech různých oborů. Výzkum je založen na dvou hypotézách:

1. Metodou automatického vyhledávání termínů pomocí data miningu založenou na kvantitativních vlastnostech lze v reálných textech rozpoznat termíny.
2. Pokud metoda automatického vyhledávání termínů rozpozná v textech termíny, bude možné na základě výsledků metody identifikovat vlastnosti typické pro termíny.

Úkolem práce je tak vedle hlavního cíle, tj. doplnění stávajících popisů termínu o kvantitativní vlastnosti založené na chování termínů v reálných textech, i nalezení úspěšné metody automatického vyhledávání termínů.

Při automatickém vyhledávání termínů a zároveň při interpretaci jeho výsledků lze plně využít výhod jazykových korpusů i korpusové metodologie. Korpusový výzkum terminologie má stejné výhody jako v jakémkoli jiném odvětví lingvistiky: 1. rozsáhlý a kvalitní materiál a 2. možnosti kvantitativního zpracování těchto dat.

Termíny, tedy „názvy (lexémy) užívané jednoznačně pro denotáty specifické určité vědě, oboru, ale i řemeslu či speciálnímu povolání“ (Čermák, 2010, s. 132), tvoří největší část slovní zásoby každého rozvinutého jazyka (tamtéž). Slovo nebo kolokace je termínem v určitém oboru, kde má místo v systému ostatních termínů a kde má jasně vymezený význam (ať už definicí, nebo např. konvencí). Výzkum termínů bývá zaměřen teoreticky[1], tj. na popis termínů a terminologie (ve smyslu soustavy termínů daného oboru), nebo prakticky[2], např. na sestavování terminologických slovníků či databází a na automatické vyhledávání termínů v textech. Představovaný výzkum spojuje obě zaměření - na základě praktického úkolu vyhledávání termínů zjišťuje teoretické poznatky o termínech.

Terminologické práce se často soustředí na to, jak by termín jako součást terminologie (či nomenklatury) měl vypadat, aby byl funkční v systému ostatních termínů v daném oboru (např. Poštolková et al., 1983). Z toho vyplývá i popis termínu založený na (žádoucích) kvalitativních vlastnostech, který je pak však jen málo použitelný pro rozlišování termínů od netermínů v reálných textech. Tento poznatek zmiňuje i Machová (1995, s. 138):

[1] Čermák, 2010; Bozděchová, 2009; Cabré, 2003; Machová, 1995; Poštolková et al., 1983; Bečka, 1972
[2] Šrajerová, 2009b; Wermter, Hahn, 2005; Chung, 2003; Frantzi, Ananiadou, 1997; Yang, 1986

„Přistoupil-li by terminograf k textu a chtěl-li by s těmito návody (definicemi a kvalitativními popisy termínů, pozn. autorky) odlišovat termíny od netermínů, je možné, že by neuspěl."

Spolehlivé rozlišení termínů od netermínů v reálných textech je obtížným a problematickým úkolem. Mnoho autorů upozorňuje na plynulý přechod mezi termíny a netermíny (Kocourek, 1965, s. 24; Bečka, 1972, s. 49; L'Homme et al., 2003, s. 154). Nejde přitom jen o některé specifické disciplíny, dokonce i v oborech obsahově nejstabilnějších, jako je např. lékařství, bývá někdy odlišení termínů a netermínů nejednoznačné (Bozděchová, 2009, s. 8). Za nejasností hranice mezi termíny a netermíny stojí škálovitá povaha termínu, jinak řečeno různě silná terminologická platnost (Čermák, 2010, s. 134)[3]. Příkladem různé terminologické platnosti můžou být termíny *monokolokabilita*, *rekurzivní algoritmus* a *benzenkarboxylát draselný*, které je v rámci příslušných oborů (lingvistika, výpočetní technika a chemie) jednodušší jednoznačně zařadit mezi termíny než např. *slovo*, *prostředí* a *ohřívat*. Škále v terminologii je věnována kapitola 1.3.

Intuitivně vnímaná síla terminologické platnosti[4] je využitelná především při ručním vyhledávání termínů. Je nutně subjektivní a z toho důvodu je i označování termínů, zvláště těch se slabou terminologickou platností, do určité míry nekonzistentní. Proto je ruční označování termínů, které je základem přípravy dat pro představovanou metodu automatického vyhledávání termínů (ATR, Automatic term recognition), založeno nejen na intuici, ale i na terminologických slovnících a na spolupráci s odborníky zkoumaných disciplín. Přípravou dat i diskuzí o problematice ručního vyhledávání termínů se zabývá kapitola 2.2.5.1.

Předkládaný výzkum zavádí jednotný postup výpočtu přesných hodnot terminologické platnosti pro všechna slova v textu, a je tedy konzistentní[5]. Výpočet konkrétních hodnot terminologické platnosti je založen na přesných hodnotách kvantitativně vyjádřených vlastností slov v textech a jeho výsledkem je rozdělení všech slov ve zkoumaných textech na termíny a netermíny. Metoda výpočtu terminologické platnosti je podrobně popsána v kapitole 2.3.

Zkoumání a zpracovávání termínů z lingvistického hlediska je velmi náročný úkol. Jedinec totiž nemůže obsáhnout všechny vědomosti a znalosti, které by bylo třeba mít pro vyčerpávající a dokonale informované zpracování terminologií všech oborů. Každá obo-

[3]Podobně by se dalo mluvit i o centru a periferii terminologie, či o prototypických termínech.
[4]Intuitivně vnímaná škála je ovlivňovaná řadou faktorů, kterými se zabývá kap. 1.3.1.
[5]Spolehlivý je ovšem jen do té míry, do jaké je spolehlivá sama metoda vyhledávání termínů.

rová terminologie má svá specifika, která znají a rozumí jim jen odborníci z daného oboru. V některých oborech se v rámci terminologie používají různé zkratky, značky nebo číslovky, např. H_2SO_4, v dalších je zase obvyklé a časté tvoření víceslovných termínů s použitím vlastních jmen (např. *Hallův teorém, Eukleidova věta*), což je v jiných oborech postup jen okrajový nebo neobvyklý (Hausenblas, 1962). Kromě toho se obor od oboru liší také terminologická a terminografická praxe, v některých disciplínách existují např. ucelené nomenklatury nebo nejrůznější normy, které v jiných oborech nejsou vůbec známy. Liší se také typy textů, které jsou v rámci oboru užívány: jen těžko bychom v lingvistice nebo v biologii hledali např. obdobu lékařské zprávy jako jedinečného způsobu komunikace mezi odborníky, nebo třeba návodu na použití nějakého elektronického přístroje.

Pro lingvistu by bylo jistě snazší studovat pouze terminologii lingvistickou, ale takový přístup k obecnějšímu studiu termínů není přijatelný. Existují dvě možnosti, jak se s tímto problémem vyrovnat: buď je možné v rámci velkého projektu zapojit do terminologického zkoumání nejen lingvisty, ale i odborníky z mnoha nejrůznějších oborů (o nutnosti spolupráce s odborníky z jednotlivých oborů viz Bozděchová, 2009, s. 29), nebo je třeba se pokusit na základě informací dostupných jedinci zpracovat terminologii co nejobsáhleji, ale s ohledem na to, že mnohé zůstane příliš zobecněno či zjednodušeno a nezpracováno. Vzhledem k rozsahu dizertační práce jsem zvolila druhou cestu.

Problémy terminologie nespočívají jen v rozsáhlosti zkoumaného předmětu nebo v nemožnosti se stoprocentní jistotou rozlišit termíny od netermínů. Se zařazováním termínu do určitého oboru totiž úzce souvisí i poměrně komplikovaná (a v rámci lingvistické práce jen těžko řešitelná) problematika stanovení hranic oboru. V terminologii se přirozeně velmi často používají slova „obor“ nebo „disciplína“, případně „tematická“ či „odborná oblast“, do které patří určitá soustava termínů. V pracech o terminologii a v definicích termínu ale není vždy jasné, co se „oborem/disciplínou/tematickou oblastí“ míní. Může jít o velmi široké i velmi úzké pojetí[6], škálu lze ukázat na čtveřici *přírodní vědy - obecná biologie - botanika* - (úzce specializovaná) *archeobotanika*. Označení „tematická oblast“ problematizuje i Machová: „Hranice tematické oblasti lze stanovit do značné míry libovolně. Současně je nikdy není možno stanovit striktně. Navíc dochází (...) k jejich neustálé změně.“ (Machová, 1995, s. 140)

Řešením problematiky stanovení hranic oboru se zabývají autoři článku shrnujícího vývoj terminologie v prvních deseti letech existence časopisu Terminology: Podle nich je nutné

[6]Obdržálek (2001) mluví v této souvislosti o různých rovinách preciznosti (čím úže změřený obor, tím musí být k terminologii přistupováno precizněji).

si uvědomit, že definice „oboru“ je arbitrární a relativní vzhledem k cílům konkrétního terminologického projektu. Je tedy třeba vždy vymezit v rámci daného výzkumu, co budeme považovat za obor (L'Homme et al., 2003, s. 153). V rámci korpusového výzkumu termínů je výhodné využít takového rozdělení, jaké poskytuje použitý korpus - např. ve zde použitém korpusu SYN2010 jsou jednotlivé obory rozděleny v rámci metainformace „genre“.

V předkládané dizertační práci byla pro výzkum vlastností termínu vyvinuta originální metoda ATR založená na data miningu (na vytěžování informací z velkého množství dat). Metoda pojmenovaná TERMIT (Term Mining Tool) umožňuje identifikovat jednoslovné i víceslovné termíny v textech nejrůznějších akademických oborů, a to s vysokou úspěšností: zhruba 95 % slov je správně zařazeno mezi termíny nebo netermíny. Nové metodě TERMIT se věnuje kapitola 2.3, přehled měření úspěšnosti metody v různých oborech lze nalézt v kapitolách 3.1 a 3.2.

Na základě výzkumu lze určit, které ze zkoumaných frekvenčních, distribučních, lingvistických a kontextových vlastností jsou charakteristické pro jednoslovné a víceslovné termíny. Tyto typické vlastnosti jsou navíc podkladem pro budoucí výzkum v oblasti ATR, protože mohou být základem dalších úspěšných metod automatického vyhledávání termínů. Pro jednoslovné termíny jsou to vlastnosti založené na porovnání frekvence slova v oboru a v neakademických textech, na distribuci slova v různých disciplínách a na rovnoměrnosti rozmístění slova v korpusu. Víceslovné termíny obvykle vykazují vysoké hodnoty asociačních měr (používaných při hledání kolokací) a typicky obsahují aspoň jeden termín jednoslovný. Identifikaci víceslovných termínů usnadňuje i jejich chování ve vzájemném kontextu. V kapitolách 5.1.1 a 5.2.1 jsou jednotlivé zkoumané vlastnosti termínů seřazeny podle důležitosti při procesu vyhledávání termínů. Popis všech vlastností použitých v rámci výzkumu poskytuje kapitola 2.2.4.

Termíny byly automaticky označeny v textech téměř 40 oborů dostupných v korpusu SYN2010, což umožňuje mj. zjistit procento jednoslovných a víceslovných termínů v textech i ve slovní zásobě konkrétních oborů, a dále mezi sebou jednotlivé disciplíny porovnávat. Přehled počtu jednoslovných termínů v několika desítkách oborů je v kapitole 4.1.2, počet víceslovných termínů ve čtyřech různých oborech v kapitole 4.4. Kapitola 4.2 se věnuje zkoumání příbuznosti mezi obory na základě vyhledaných termínů a porovnání humanitních a přírodovědných/technických oborů.

V neposlední řadě přináší práce poznatky cenné pro korpusovou metodologii, konkrétně zhodnocení úlohy lingvistického značkování (lemmatizace a morfologické tagování) v tomto typu výzkumu (kap. 2.2.6).

1 Teoretická východiska

Kapitola zachycující teoretická východiska celé práce se věnuje dosavadnímu výzkumu v oblasti terminologie a přínosu automatického vyhledávání termínů a korpusové lingvistiky pro zkoumání a popis termínů. Vysvětluje, jaké jsou meze terminologického zkoumání, ať už jde o množství zkoumaných oborů, z nichž každý je nějakým způsobem specifický, nebo o problémy sahající ještě hlouběji, jako je problematické stanovení hranic samotných oborů.

Druhá část kapitoly je zaměřena na existující definice termínu a na popis jeho jednotlivých vlastností, kvalitativních i kvantitativních. Zmiňuje především takové vlastnosti, které můžou být využity při automatickém vyhledávání termínů, tedy takové, které lze zpracovat za pomoci počítačových nástrojů. Mezi hlavní charakteristiky, jejichž role v rámci termínů bude dále zkoumána, patří frekvence a distribuce v akademických i neakademických textech, kolokační chování a chování v kontextu.

Závěr kapitoly se zabývá principem škály v terminologii, tedy tím, že některé termíny jsou vnímány jako silnější a jiné jako slabší. Na hodnotu terminologické platnosti slov nebo kolokací působí různé vlivy, vnější i vnitřní. Mezi vnější vlivy patří odbornost textu, v kterém se dané slovo či kolokace vyskytuje, a zároveň příslušnost k oboru, jehož je daný text součástí. Terminologickou platnost ovlivňuje i případná polysémie daného slova (ač tradičně popisovanou vlastností termínu je právě jeho jednoznačnost). Kapitola je zakončena konstatováním, že právě přijetí principu škály je zásadní pro automatické vyhledávání termínů, zvláště pro interpretaci jeho výsledků.

1.1 Terminologie a korpusový výzkum termínů

1.1.1 Termín a slovo: terminologie v rámci lingvistiky

Terminologie se soustředí především na: 1. výzkum terminologie (ve smyslu úhrnu termínů) či nomenklatur[1] konkrétních oborů a výzkum základní terminologické jednotky - termínu, 2. vytváření databází termínů a terminologických slovníků, sloužících jak lokální, tak mezinárodní odborné komunitě, a 3. standardizaci/normalizaci termínů a terminologií. Stan-

[1] V některých oborech (zvl. v přírodních vědách a v technických oborech, které se ne náhodou o terminologii zajímaly dříve než ostatní obory, vznikly závazné názvoslovné normy, jejichž výsledkem je nomenklatura (názvosloví). Nomenklatura je systém určité části termínů v rámci oboru, který zahrnuje nejen samotné termíny, ale i klasifikaci jevů, příp. jejich uspořádání do hierarchického systému (Čermák, 2010, s. 133).

dardizace má v terminologii pevné místo, má totiž zásadní význam pro porozumění mezi odborníky dané disciplíny (zvláště pokud jde o obory technické)[2].

Klasifikace termínů vychází z různých kritérií, časté je např. dělení termínů podle oboru, do kterého náležejí, což může být v některých případech problematické[3], nebo pro tuto práci (a pro ATR vůbec) velmi důležité rozlišení termínů jednoslovných a víceslovných. Někdy jsou z terminologie vydělovány tzv. profesionalismy a slova slangová - u různých autorů jsou taková slova či kolokace chápány různě, a to jak z hlediska významu, konkrétně expresivnosti, tak z hlediska spisovnosti (pro přehled přístupů k těmto výrazům viz Bozděchová, 2009, s. 36). Čermák (2010, s. 135) toto umělé rozdělování termínů na základě spisovnosti podrobuje kritice: to, že je výraz používán často neformálně a že není považován za spisovný, ještě neznamená, že má nižší hodnotu a je méně vhodný a že by neměl být zahrnut do terminologie daného oboru.

Terminologie[4] je poměrně mladá disciplína, počátky terminologie jako oboru se datují do 30. let 20. století[5]. Tehdy svými myšlenkami výrazně zapůsobil zakladatel moderní terminologie, rakouský terminolog E. Wüster. Především od 50. do 80. let 20. století trvalo poměrně stabilní období rozvoje a rozšiřování oboru, během nějž nedocházelo ke zpochybňování základů terminologie, jejích jednotek nebo cílů. V 90. letech 20. století se ale v rámci disciplíny rozpoutala diskuze o přehodnocení tradiční teorie terminologie. Šlo především o terminologickou jednotku - termínu, zvláště pak jeho vztah k lexému, a o postavení terminologie jako samostatné disciplíny (Cabré, 2003, s. 169-171)[6].

Jednou z hlavních otázek současné terminologie je vztah termínu a lexému (slova či kolokace) a s tím související vztah terminologie a lingvistiky, konkrétně zda je termín zvláštním případem slova, nebo je to zcela odlišná jednotka, a zda je terminologie samostatnou disciplínou, nebo součástí lingvistiky[7]. Podle některých autorů se termíny z lingvistického

[2]Ke standardizaci z pohledu automatického vyhledávání termínů viz kap. 1.2.2.2.

[3]Např. termín *kyslík* patří primárně do chemie, ale pokud je v rámci popisu fotosyntézy zmiňován v botanických textech, nelze ho vyřadit jako termín nenáležející (alespoň sekundárně) do oboru botaniky.

[4]*Terminologie* není jen název oboru, ale i název pro jeden z předmětů terminologického zkoumání, soubor termínů v určitém oboru, nebo dokonce soubor termínů v jazyce obecně. Termín *terminologie* má tedy hned tři různé významy. Jedním z často zmiňovaných požadavků na termín přitom je, že by v ideálním případě neměl být polysémní (Čermák, 2010, Poštolková et al., 1983, Bozděchová, 2009). Už samotný název disciplíny tedy ukazuje rozdíl mezi ideálem a realitou.

[5]Zájem o terminologii se ale projevoval mnohem dříve, už od 18. století, především v oborech s velmi rozsáhlým názvoslovím: v chemii, v botanice a v zoologii. Na začátku 20. století přebraly díky velmi rychlému rozvoji vůdčí roli v terminologii obory technické (Cabré, 1999, s. 1-2).

[6]Více o historii disciplíny Bozděchová (2009, s. 18-21).

[7]Velmi podrobně a přehledně zpracovala tyto otázky Cabré (1999).

hlediska chovají stejně jako neterminologické lexikální jednotky (L'Homme et al., 2003, s. 154) a jejich zkoumání patří z pohledu lingvistiky do lexikologie. Další naopak upozorňují na specifičnost terminologie v rámci lexikonu spojenou s významem termínů. Ten totiž často bývá v rámci daného oboru přesně definovaný, nebo ustálený (např. má pevné místo v nomenklatuře nějakého oboru), a tím se liší od slov obecného jazyka, která významu nabývají až v kontextu (Palmer, 1968, s. 179). Čermáková a Teubert, kteří vidí jako hlavní úkol korpusové lingvistiky zkoumání změn významu slov v různých kontextech, na tento fakt upozorňují: Jakmile je termín v rámci nějakého oboru přesně definován a zařazen do systému termínů tohoto oboru, přestane být přirozeným slovem jazyka a z hlediska významu se začne chovat nestandardně - jeho význam není závislý na kontextu (Teubert, Čermáková, 2004, s. 141).

Zajímavý je pohled na termíny z hlediska lexikografie: termíny jsou ve slovnících zpracovávány různým způsobem, ať z hlediska výběru hesel pro slovník nebo z hlediska samotného tvoření slovníkového hesla (jasné rozdělení neterminologického a terminologického výkladu). V českém prostředí existují např. slovníky cizích slov, což jsou v podstatě slovníky (obecně) terminologické, které zahrnují jak slova vyskytující se napříč všemi disciplínami (to jsou např. termíny z oboru teorie vědy či termíny statistické, viz kap. 1.3.1.2), tak specializované termíny z jednotlivých disciplín.

Praktičtějším řešením by bylo při vytváření kvalitního výkladového slovníku zahrnout do hesláře právě i tyto termíny, ovšem vždy s jasným označením, že se jedná o termín (u polysémních slov s významem terminologickým i neterminologickým by heslo mělo být podle toho rozděleno). Heslář takového velkého výkladového slovníku by měl být vytvářen podle frekvence slov v textech (v korpusu jazyka). U termínů by samozřejmě bylo třeba zvolit ještě další kritéria, jako je například distribuce v dalších oborech a v obecných textech, případně rozložení v korpusu (viz kap. 1.2.2.1) - tím by bylo možné vyloučit vysoce specializované termíny, jejichž vysoká frekvence v korpusu souvisí pouze s vysokou frekvencí např. v jediném textu v rámci jediného oboru. Nejzajímavější z hlediska lexikografie jsou termíny polysémní, zvláště pak ty, které pronikají z odborných textů do textů obecných (publicistika, beletrie, neformální mluvené projevy atd.).

1.1.2 Automatické vyhledávání termínů

Možnosti automatického zpracování jazyka jsou v současné době díky vyspělosti informačních technologií a díky obrovskému množství kvalitního lingvistického materiálu

(jazykové korpusy) omezené v podstatě pouze tím, jak jsme schopni nabízených zdrojů využít. V posledních dvaceti letech se zájem komputační lingvistiky zaměřil i na terminologii, konkrétně na automatické vyhledávání (rozpoznávání) termínů v textech. Desítky metod využívají nejrůznější vlastnosti termínů a jejich kontextu k tomu, aby co nejspolehlivěji našly co největší množství termínů v textech (viz kap. 1.2.2.1).

Efektivní metoda automatického vyhledávání termínů je základem pro mnoho aplikací, jako je překlad nebo strojový překlad, tvorba terminologických databází a terminologických slovníků jedno- i vícejazyčných, automatická indexace textů nebo charakterizace textových typů (Yang, 1986; Kageura, Umino, 1996; Chung, 2003). Kageura a Umino (1996, s. 18) ale poukazují na to, že metody automatického vyhledávání termínů založené na statistických rysech mají nejen praktické využití, ale jejich výsledky mají dopad na teorii terminologie (popis vlastností termínu). Pokud zjistíme, které vlastnosti hrají v automatickém vyhledávání termínů nejdůležitější roli, můžeme pomocí nich charakterizovat termín nebo jeho popis smysluplně doplnit. K tomu je ovšem třeba, aby daná metoda 1. byla poměrně úspěšná co do kvality a kvantity vyhledaných termínů a aby 2. byla schopna určit, které ze zkoumaných vlastností termínu byly pro vyhledání skutečně podstatné (k tomu viz kapitoly o seřazení vlastností termínu podle důležitosti, 5.1.1 a 5.2.1).

O zájmu o metody ATR svědčí velké množství přehledových článků, hodnotících jednotlivé metody z hlediska úspěšnosti, využitelnosti i z hlediska použitých lingvistických, statistických či dalších rysů (Kageura, Umino, 1996; Chung 2003; Kageura et al., 2004; Lemay et al., 2005; Kit, Liu, 2008; Vivaldi et al., 2012; Marín, 2014).

Některé metody ATR využívají především kvantitativních (statistických) rysů, především frekvenci slova v textech či distribuci v korpusu textů. Yang (1986) zakládá svou metodu na předpokladu, že konkrétní termíny typicky přináležejí do konkrétní disciplíny[8]. K vyhledávání jednoslovných termínů využívá frekvenci a distribuci a k vyhledání víceslovných termínů navíc asociační míry (používané pro identifikaci kolokací). Podobně Kageura a Umino (1996) zavádějí pojem termhood, který vyjadřuje míru přináležitosti termínu do určitého oboru (jde o specifický přístup k distribuci). Chung (2003) také využívá v jednoduché metodě ATR frekvenci a počty dokumentů obsahujících zkoumané slovo (tedy zvláštní případ distribuce). Na distribuci je založena i metoda vyhledávání víceslovných termínů Wermtera a Hahna (2005). Kit a Liu (2008) využívají pouze frekvenčních charakteristik slov a řadí je podle hodnoty poměru frekvence výskytů v určitém oboru ku

[8]V anglickém originále: „(...)terms are highly subject matter specific“ (Yang, 1986).

srovnávacímu korpusu.

Jiné metody používají k vyhledávání termínů lingvistické vlastnosti, jako jsou morfologická či syntaktická specifika (Ville-Ometz et al., 2007; Nenadić et al., 2004), či seznamy slov, která určitě nemohou být termínem nebo která se často vyskytují/vůbec nevyskytují v okolí termínu apod. (mezi takové metody se řadí i Frantzi, Ananiadou, 1996, 1997 a 1999).

1.1.3 Korpusová lingvistika a korpusový výzkum terminologie

Jazykový korpus je rozsáhlý soubor elektronicky uložených a zpracovávaných jazykových dat (Čermák, 2000, s. 15), který slouží primárně k jazykovému výzkumu. Zkoumáním tohoto materiálu specifickými metodami se zabývá korpusová lingvistika. Z pohledu terminologického bádání jsou v českém prostředí nejlépe využitelné reprezentativní synchronní korpusy psaných textů, především aktuální SYN2010, a to díky tomu, že ve vyváženém poměru obsahuje texty jak odborné[9], tak i publicistické a beletristické, které mohou tvořit srovnávací bázi pro sledování specifičnosti chování termínů.

Korpusová lingvistika je empirická disciplína, která pracuje s rozsáhlými autentickými daty, jež jsou součástí skutečného úzu: v případě zde použitého korpusu SYN2010 jde o reálné publikované texty různých typů. Používá specifické metody práce, jako je např. statistické zpracování dat, a přináší mnoho nových poznatků o jazyce. Teubert (2005b, s. 7-8) upozorňuje na to, že v korpusové lingvistice jakožto oboru humanitním nejde o hodnocení toho, co je správné nebo přípustné, ale o interpretaci výsledků analýzy a o jejich předložení jazykovému společenství.

Ke zkoumání korpusových dat můžeme přistupovat dvěma způsoby: buď na základě introspekce předem vytvoříme hypotézu, pro jejíž potvrzení (resp. vyvrácení) hledáme v korpusu argumenty (corpus-based přístup), nebo popisujeme jazykové jevy až v závislosti na výsledcích analýzy dat (corpus-driven přístup)[10]. Tyto dva přístupy ve své knize rozlišila Tognini-Bonelli. Uvádí, že corpus-based přístup je zaměřený na následující úkoly: vysvětlit, testovat nebo příklady doložit teorie a popisy jazyka, které byly formulovány na základě lingvistické introspekce před vznikem velkých korpusů (Tognini-Bonelli, 2001, s. 65).

[9] Termíny se ale vyskytují nejen v textech odborných, ale v podstatě v jakémkoli typu textů (viz kap. 1.3.1.1), proto je možné terminologii studovat (podle výzkumných cílů) i v ostatních součástech Českého národního korpusu (např. i v mluvených korpusech ORAL).

[10] Termíny corpus-based a corpus-driven přístup, které se obvykle používají v původní podobě, by bylo možné přeložit jako přístup na korpusu založený, resp. korpusem ověřovaný, a přístup korpusem řízený (Čermáková, 2009, s. 21).

V corpus-driven přístupu není korpus vnímán jen jako zdroj příkladů nebo argumentů pro teorie definované na základě introspekce; teoretická tvrzení jsou podle ní odrazem korpusových dokladů a jsou s nimi zcela konzistentní. K datům se přistupuje jako k celku a zkoumají se opakující se schémata a frekvence jevů (Tognini-Bonelli, 2001, s. 84)[11]. Tento přístup umožňuje i objevy zcela nových jevů, jak tomu bylo třeba v případě kolokací (viz níže).

Mezi oběma přístupy ale ve skutečnosti v současném korpusovém výzkumu nevede takto původně zachycovaná ostrá hranice (Čermáková, 2009, s. 21). Corpus-driven přístup je spíše ideálem, ke kterému výzkumy směřují, než reálnou praxí. Pro badatele totiž není možné vystoupit z jazykového společenství a zkoumat jazykové jevy zcela nezávisle na své zkušenosti, stejně tak jako není možné oprostit se od vědomostí, znalostí a představ o jazyce nabytých v rámci lingvistického vzdělání (což by vyžadoval striktně corpus-driven přístup). Každý výzkum je založen z velké části na dosavadním poznání jazyka a na stejném základě jsou také výsledky analýzy dat interpretovány (Tognini-Bonelli, 2001, s. 85). Místo úplného zavržení teoretických východisek je realističtější usilovat o oslabení jejich role: „Analýza korpusových dat by tak měla ideálně vycházet z minimálních teoretických východisek a hypotézy založené na pozorování dat“ (Čermáková, 2009, s. 23). Badatel by v každém případě měl být připraven přehodnotit obecně známé a přijímané způsoby popisu jazyka, pokud jim korpusová data neodpovídají.

Pro terminologii má význam i ústřední pojem korpusové lingvistiky: kolokace[12]. Velkou část termínů totiž tvoří termíny víceslovné, které se řadí mezi systémové pravidelné kolokace (Čermák, 2010, s. 214). Vyhledávání kolokací v korpusovém materiálu je běžnou záležitostí, slouží k němu nejrůznější lexikální asociační míry (Pecina, 2009; Cvrček, 2013). Právě lexikální asociační míry mohou velmi přispět k úspěšnosti automatického vyhledávání (víceslovných) termínů.

Významnou roli ve výzkumu terminologie (stejně jako ve výzkumu jazyka obecně) hraje

[11]Sinclair (2004, s. 190-191) k tomu dodává, že v corpus-driven lingvistice je nutné omezit používání lingvistického značkování, protože to už v sobě zahrnuje lingvistickou interpretaci - je třeba pracovat s čistým textem, v kterém se vyhledávají opakující se vzory či schémata. Praktickou výhodou, která vyplývá z omezení značkování, je možnost pracovat s texty netagovanými nebo nelemmatizovanými (tedy nejen s těmi, které jsou k dispozici v korpusech).

[12]Kolokace je v českém prostředí definována jako „(smysluplné) spojení lexémů/slov, lexikální syntagma, zvl. v podobě víceslovného pojmenování, jehož vznik je podmíněn vzájemnou kolokabilitou, a tedy i kompatibilitou“ (Čermák, 2001, s. 254). John Sinclair kolokaci definuje odlišně, obecněji, jako výskyt dvou nebo více slov v textu blízko sebe - zajímavé pro lingvistiku jsou především ty kolokace, které se často opakují (oproti neočekávaným (dramatickým) kolokacím) (Sinclair, 1991, s. 254).

frekvence slov a jazykových jevů, která se za pomoci jazykových korpusů dá zkoumat daleko jednodušeji a spolehlivěji než kdy předtím. Frekvencí v jazyce se zabýval už v r. 1935 americký lingvista G. K. Zipf (1935) a po něm řada dalších lingvistů, v českém prostředí např. M. Těšitelová nebo J. Králík, J. V. Bečka dokonce z pohledu terminologie (Bečka, 1972), a dále autoři nejnovějších frekvenčních slovníků češtiny (Čermák, Křen, 2004; Čermák, Křen, 2011). Pro automatické vyhledávání termínů má frekvenční analýza zcela zásadní význam, což naznačuje i množství metod ATR využívajících frekvenční charakteristiky slov k identifikaci termínů. Zjednodušeně se dá říct, že termíny jsou výrazně a přirozeně frekventovanější v odborných než např. v publicistických či beletristických textech (Machová, 1995, s. 139). Už zcela jednoduchou metodou založenou právě na poměru těchto frekvencí by bylo možné vyhledat velké procento termínů v akademických (a jiných odborných) textech (takovou jednoduchou metodu představuje Kit, Liu, 2008).

1.2 Definice a vlastnosti termínu v dosavadních výzkumech

1.2.1 Definice termínu

V odborné literatuře je možné najít desítky definic termínu a mnoho komentářů k nim[13]. Nejčastější definicí termínu, se kterou se setkáme ve studiích o automatickém vyhledávání termínů, je zachycen v normě ISO 1087:1990, kde je termín definován jako „designation of a defined concept in a special language by a linguistic expression“[14], což můžeme přeložit jako „označení definovaného pojmu v odborném jazyce lingvistickým výrazem“.

Česká odborná literatura věnuje termínu a jeho popisu velkou pozornost. Výchozí definicí pro tuto práci je už výše zmiňovaná definice Čermákova[15]. Je výstižná a přitom dostatečně obecná a neomezuje termín požadavky např. na spisovnost či definovanost, které jsou obsaženy v některých dalších popisech termínu.

Podobně obecně přistupuje k termínu Machač (1964). Termín od neterminologických pojmenovacích jednotek podle něj odlišuje „specificky vymezený význam a specifičnost funkce

[13]Přehledné porovnání nejrůznějších definic termínu v české odborné literatuře lze najít např. v pracech Kocourka (1965) nebo Bozděchové (2009).

[14]Jde o starší práce před rokem 2000 (např. Lauriston, 1995). Aktuální definice termínu je součástí normy ISO 1087-1:2000: „Verbal designation of a general concept in a specific subject field.“ s doplněním: „A term may contain symbols and can have variants, e.g. different forms of spelling.“

[15]Termíny jsou „názvy (lexémy) užívané jednoznačně pro denotáty specifické určité vědě, oboru, ale i řemeslu či speciálnímu povolání“ (Čermák, 2010, s. 132).

v dorozumívacím styku s omezením na příslušný obor vědecký nebo praktický“ (Machač, 1964, citováno dle Kocourek, 1965, s. 15). Všechny ostatní vlastnosti, které jsou termínu připisovány, jsou „více nebo méně závažné a typické jen za určitých okolností“ (tamtéž). Mezi nejproblematičtější vlastnosti, požadované některými definicemi termínu (viz níže), patří definovanost termínu a jeho jednoznačnost.

1.2.1.1 Definovanost termínu

Definovanost jako zásadní vlastnost termínu požaduje např. Kocourek. Ten vymezuje termín velmi jednoduše, jako „definované slovo nebo sousloví“ (Kocourek, 1965, s. 21). Taková definice zachycuje velkou část termínů, zvláště v technických, přírodovědných a formálních vědách. Na druhou stranu mnoho slov, která bychom mezi termíny jednoznačně zařadili, by na základě tohoto požadavku bylo vyřazeno, zvláště pak nejrůznější ad hoc termíny často používané v humanitních vědách[16]. V humanitních oborech je také běžné, že jediný termín je chápán různě a má i mnoho různých definic (příklady jen z oboru lingvistika mohou být *slovo*, *diskurz*, nebo dokonce i *termín*).

Užitečné z tohoto pohledu tedy může být vyčlenění dvou druhů terminologií, jak to činí např. Machová (1995, s. 143-144). Machová rozeznává terminologie preskriptivní, kde převažuje typ termínů, které jsou přesně definovány při prvním použití a dále jsou užívány pouze v tomto definovaném významu, a terminologie pseudopreskriptivní, kde je užití termínu dáno subjektivním postojem či náhledem konkrétního člověka (či školy apod.), příp. je význam termínu vymezen pouze dohodou akceptovanou širším společenstvím. K prvnímu druhu patří terminologie formálních a přírodních věd, stejně jako terminologie technických oborů. K terminologiím pseudopreskriptivním patří většina humanitních věd.

Podobně Teubert a Čermáková (2004) rozlišují mezi terminologiemi tvrdými a měkkými. Tvrdé terminologie jsou (v rámci možností) přesné a neměnné, kdežto měkké jsou dynamické, v jejich rámci vznikají nové termíny, které nejsou přesně definovány, nebo jsou definovány různě v různých pracech podle osobního názoru daného autora.

Některé popisy termínů relativizují požadavek na definovanost i v jiných než humanitních oborech: za termíny považují i lexémy, které nejsou definované, ale jejichž přesný význam

[16]Bečka poukazuje na to, že vedle plně etablovaných termínů existují i tzv. „working terms“, pracovní termíny, použité ad hoc v určitém kontextu nebo textu (Bečka, 1972, s. 48).

je dán konvencí - např. v technických oborech se určité součástky nedefinují, přesto jsou jejich názvy termíny (Filipec, Čermák, 1985, s. 95).

1.2.1.2 Jednoznačnost termínu

Další vlastností vyžadovanou některými definicemi je jednoznačnost termínu. Kopeckij (1935) definuje termín jako „slovo, které má v odborném jazyku přesný a jednoznačný význam a které, i když se vyskytne v jazyku hovorovém, je pociťováno jako slovo náležející k některé odborné oblasti“ (Kopeckij, 1935, citováno dle Kocourek, 1965, s. 5). Stejně jako v případě definovanosti, i tato podmínka platí pro mnoho, dokonce většinu termínů, nikoli však pro všechny. Slovo *sůl*[17] v běžné komunikaci (*Podej mi prosím tu sůl.*) není pociťováno jako termín náležející do oblasti chemie, přesto v této disciplíně je jednoznačně termínem. To platí pro mnohé další předměty denní potřeby, které jsou v nějakém oboru nebo řemesle termínem[18].

Podobně problematický je požadavek, aby termín náležel do jediného oboru. Podle Havránka (2003, citováno dle Bozděchové, 2009, s. 32) jsou termíny „jednoznačná slova, kterých se užívá v jediném oboru a která v základě podržují svůj odborný význam, i vyskytnou-li se v řeči o oboru jiném nebo v běžném jazyce“. Slovo *komunikace*[19] je přitom termínem v lingvistice, dopravě i biologii a ve všech případech má rozdílný význam (i když s jistým společným základem, jímž je transport informace/dopravního prostředku/živin z jednoho místa do druhého). Více k polyfunkčnosti či mezioborové polysémii v kap. 1.3.1.5.

1.2.1.3 Definice založená na kvantitativních rysech

Stávající definice termínu vycházející především z kvalitativních lingvistických vlastností mohou být na základě výzkumu doplněny charakteristikami kvantitativními. Stejně jako u většiny vlastností kvalitativních je ale třeba počítat s tím, že nebudou spolehlivě platit pro všechny termíny[20].

[17] Příklad slova *sůl* je vypůjčen od F. Čermáka, osobní komunikace.

[18] Podle Čermáka se to týká většiny nejfrekventovanějších konkrét (Čermák, 2010, s. 134).

[19] Dalším příkladem může být termín *čípek* v lékařství a botanice, přičemž v lékařství má dokonce několik významů: léčivý přípravek specifických vlastností a tvaru, buňka v oční sítnici, *čípek patrový* a *čípek děložní* (což jsou však jednoznačné víceslovné termíny).

[20] Vždy existují výjimky, a často právě výjimečná slova nebo termíny mají vysokou frekvenci (čím frekventovanější slovo, tím je pravděpodobnější, že bude vykazovat nepravidelnosti a odchylky od běžného chování).

Bečka se pokusil o popis termínu na základě kvantitativních vlastností; charakterizuje slova s terminologickou platností „jako jednotky, které mají nízkou celkovou (absolutní) disponibilitu, t.j. jako slova s podprůměrnou celkovou frekvencí a nízkou textovou i oborovou frekvencí.“[21] (Bečka, 1972, s. 53). Na příkladu Bečkova popisu termínů se projevuje úskalí popisu termínů pomocí kvantitativních rysů - je totiž zapotřebí jednotlivé vlastnosti i jejich hodnoty formulovat dostatečně obecně, aby bylo možné je připsat co největšímu počtu termínů, a srozumitelně, aby byly uchopitelné v rámci lingvistického výzkumu.

1.2.2 Vlastnosti termínů využitelné v ATR

Odborná literatura se při popisu termínů často zaměřuje na kvalitativní charakteristiky, jako je jednoznačnost, definovanost, ustálenost, systematičnost (termín zapadá do systému ostatních termínů v dané oblasti), významová přesnost, nemetaforičnost nebo nepřítomnost expresivity (Poštolková et al., 1983; Bozděchová, 2009; Čermák, 2010). Mnoho z těchto vlastností však není využitelných při vyhledávání termínů v textech, ať už ručním, nebo automatickém.

Tento postřeh učinil už v 70. letech J. V. Bečka, podle nějž neexistuje žádné kvalitativní kritérium, na jehož základě by se v textu spolehlivě a jednoznačně vydělily termíny od netermínů[22] (Bečka, 1972, s. 50). Podobně i Machová (1995) je skeptická ohledně využití kvalitativních vlastností při analýze reálných odborných textů (viz citaci na str. 1-2).

V této kapitole je větší prostor poskytnut pouze těm vlastnostem termínů, které mají význam pro automatické vyhledávání termínů v textu[23]. Jde primárně o kvantitativní vlastnosti založené většinou na frekvenci nebo distribuci slov v textech, případně na kolokačních vlastnostech a kontextu, ale i o délku termínů a netermínů (kap. 1.2.2.1).

Zařazeny jsou i některé vlastnosti kvalitativní (kap. 1.2.2.2), které je možno v daném výzkumu využít, tj. ty, které lze vyhodnotit automaticky a které se dají vyjádřit způsobem vhodným pro automatické zpracování textu (neobvyklost formy, zařazení do slovního

[21] Vysvětlení pojmů viz níže (kap. 1.2.2.1). V angl. originále: „words with terminological validity may as lexical components be characterized in quantitative terms, viz. as items displaying low total (absolute) disponibility, i. e. as words with a total frequency well under the average, as well as low textual and register frequencies.”(Bečka, 1972, s. 53)

[22] V anglickém originále: „...there is no qualitative criterion which would delimit, in a reliable and unambiguous way (...) terms from non-terms.”(Bečka, 1972, s. 50)

[23] O tom, zda určitou vlastnost zařadit, se rozhoduje na základě předchozích výzkumů věnovaných automatickému vyhledávání termínů (především Yang, 1986; Chung, 2003; Kageura et al., 2004; Šrajerová 2009a,b; Šrajerová et al., 2009).

druhu, malé či velké písmeno na začátku slova). Pozornost je věnována také těm vlastnostem, které hrají roli při přípravě materiálu (definovanost, ustálenost) či které se promítají do frekvenčního, distribučního a kolokačního chování termínů (jednoznačnost).

Konkrétním vlastnostem použitým v předkládané metodě automatického vyhledávání termínů TERMIT je věnována kapitola 2.2.4.

1.2.2.1 Kvantitativní vlastnosti termínů

O zásadním významu kvantitativních rysů termínu mluví kromě výše zmíněného Bečky (1972) i H. Yang, autor prvního pokusu o automatické vyhledání termínů na základě korpusu textů (Yang, 1986). Podle něj je možné vědecké/technické termíny identifikovat na základě jejich statistického chování[24] (Yang, 1986, s. 97). I podle tohoto autora jsou nejdůležitějšími vlastnostmi jednoslovných termínů frekvence a distribuce (a další rysy odvozené od těchto dvou vlastností); u víceslovných termínů je pak rozhodující jejich kolokační chování (Yang, 1986, s. 93).

Také ve shrnutí vývoje terminologie za léta 1994-2004 (v rámci časopisu Terminology) uvádějí autoři (L'Homme et al., 2003, s. 154), že studie o automatickém vyhledávání termínů (např. články K. Kageury vydané právě v Terminology: Kageura, Umino, 1996; Kageura, 1997) dokážou rozlišit mezi termíny a netermíny právě na základě statistických charakteristik.

Hlavními kvantitativními vlastnostmi, na jejichž základě je možné oddělit termíny od netermínů v reálných textech, jsou **frekvence** a **distribuce** a z nich odvozené údaje. Ve vyhledávání víceslovných termínů navíc hrají významnou roli lexikální asociační míry, které jsou zaměřené na **kolokační chování** víceslovných termínů, a **variabilita kontextu** zkoumaných slov.

Frekvence

Absolutní frekvence slova je vodítkem pro nalezení hranice mezi termíny a netermíny pro Bečku (1972, s. 52). Podle něj mají terminologické jednotky obecně nižší frekvenci než neterminologické - se snižující se frekvencí slov vzrůstá procento termínů. Většina termínů, které mají vyšší frekvenci, je polysémních nebo polyfunkčních (mají různou funkci v různých oborech). Yang (1986, s. 98) ve své metodě používá tzv. **průměrnou frekvenci**

[24] V anglickém originále: „It is (...) possible to identify scientific/technical terms solely on the basis of their statistical behaviour" (Yang, 1986, s. 97).

(tj. **relativní frekvenci**) - vztahuje absolutní frekvenci k počtu slov v textu. To je důležité zejména v případě, že pracujeme s množstvím různě rozsáhlých textů z různých oborů.

Další autoři pak považují za výhodné zjistit **poměr relativní frekvence v odborném textu a relativní frekvence v textu obecném** (Chung, 2003, s. 222; Kageura, Umino, 1996, s. 16)[25]. Toho, že frekvence termínu v odborném textu je obvykle vyšší než v textu obecném, příp. se termín v obecném textu nevyskytuje vůbec, si povšimla i Machová (1995, s. 139)[26].

Relativní směrodatná odchylka (Yang, 1986, s. 98) je také rysem založeným na frekvenci - tento údaj uvádí, jak je frekvence nevyrovnaná v jednotlivých oborech a v textu určeném ke srovnání (obv. beletrie či publicistika). Je to charakteristika korelující s distribucí, její výhodou je, že poskytuje o termínu podrobnější informace - ukazuje, do jaké míry kolísá frekvence mezi jednotlivými obory navzájem a také ve srovnání s obecným korpusem. U termínů bývají hodnoty značně rozptýlené.

Konkrétní rysy vycházející z frekvence použité v metodě TERMIT jsou relativní frekvence v disciplíně a v korpusu akademických a neakademických textů, a dále poměr relativní frekvence v disciplíně a v korpusu akademických textů ku relativní frekvenci v textech neakademických (rysy se značkami RFQ_{disc}, RFQ_{sci}, RFQ_{compar}, $\mathrm{RFQ}_{disc}\mathrm{RFQ}_{compar}$, $\mathrm{RFQ}_{sci}\mathrm{RFQ}_{compar}$, viz kap. 2.2.4).

Distribuce

Distribuční chování termínů je odlišné od chování neterminologických jednotek. Termíny jsou obvykle pevně zakotveny v nějakém oboru, příp. v širší tématické oblasti. Nejčastěji se vyskytují jen v jednom oboru, případně v několika oborech příbuzných[27]. **Distribuce** slova v oborech je tedy poměrně spolehlivým vodítkem při vyhledávání termínů[28], zvlášť máme-li zároveň k dispozici i informace o frekvenci. Většinu termínů navíc nenalezneme v obecných textech (zvl. termíny vysoce specializované), proto je užitečné zjišťovat i distribuci slova v odborných i neodborných textech.

[25] V podstatě totéž ale dělá i Yang (1986), který srovnává deset stejně dlouhých textů - devět odborných textů z různých oborů a jeden text beletristický.

[26] Podle Machové terminografové vědí, „zda se nějaké pojmenování v textech jejich oboru vyskytuje často/častěji, zatímco v textech neodborných se nevyskytuje/vyskytuje velmi zřídka. Jsou-li odpovědi kladné, pokládají takové pojmenování za termín a za kandidáta na zařazení do terminologického slovníku.“ (Machová, 1995, s. 139)

[27] Výjimkou jsou polysémní termíny, které se objevují ve dvou nebo více tématických oblastech, vždy s různým významem.

[28] K tomu např. Bečka (1972, s. 53), Yang (1986, s. 98) nebo Chung (2003, s. 222).

Distribuce se může týkat nejen rozmístění slov v oborech, ale i v jednotlivých textech (nebo může jít o kombinaci obojího). Chung (2003, s. 230) uvádí, že pokud se slovo vyskytuje pouze v jednom textu v rámci jednoho oboru, půjde pravděpodobně o termín. Odkazuje přitom na Bečkovy rysy **textová** a **oborová frekvence**[29]: jde o počet textů, resp. počet oborů, ve kterých se slovo objevilo (Bečka, 1972, s. 53-54)[30]. Čím nižší je textová a oborová frekvence (nebo lépe distribuce), tím vyšší je pravděpodobnost, že se jedná o termín.

Bečka oba údaje používá k výpočtu tzv. relativní disponibility, což je vztah textové a oborové frekvence k absolutní frekvenci slova v textech[31]. Zachycuje tu důležitou skutečnost, že totiž až kombinace několika rysů poskytuje informaci o terminologické platnosti zkoumaných slov (Bečka, 1972, s. 53).

Distribuci v celém korpusu, tedy nejen v rámci akademických textů, lze sledovat i na rozložení slova v korpusu, pokud je daná informace k dispozici - v ČNK vyjadřuje míru rovnoměrnosti rozložení v korpusu **průměrná redukovaná četnost** (ARF - average reduced frequency) (Savický, Hlaváčová, 2003). Vysoká hodnota ARF znamená, že slovo je v korpusu rozloženo rovnoměrně, tedy že se vyskytuje ve velkém počtu textů napříč textovými typy beletrie, publicistika a odborná literatura[32]. Termíny přitom ve většině případů bývají omezeny na malý počet oborů, nebývají tedy v korpusu rozloženy rovnoměrně.

S otázkou distribuce úzce souvisí i otázka polysémie a polyfunkčnosti (mezioborové polysémie)[33].

Konkrétní vlastnosti vycházející z frekvence použité v metodě TERMIT jsou nulový výskyt v korpusu neakademických textů, relativní distribuce v disciplínách, standardní odchylka relativních frekvencí v disciplínách, průměrná redukovaná četnost, relativní průměrná redukovaná četnost a směrodatná odchylka relativní vzdálenosti sousedních výskytů (rysy se značkami NoCompar, RDist, SDRFQ, ARF, RARF, SDRD, viz kap. 2.2.4).

Kolokační chování

Víceslovné termíny jsou pravidelné systémové kolokace (Čermák, 2010, s. 314). To zna-

[29] Původní Bečkovy termíny jsou textual a register frequency (Bečka, 1972, s. 53).

[30] Název frekvence je zde zavádějící, protože ve skutečnosti jde o distribuci.

[31] Termíny přitom mají nízkou relativní disponibilitu, to znamená, že buď mají nízkou absolutní frekvenci, nebo, pokud je jejich frekvence vyšší, se vyskytují v menším počtu textů a oborů, než netermíny. Nejvyšší relativní disponibilitu mají gramatická slova, která jsou obecně velmi frekventovaná a zároveň se vyskytují napříč všemi texty a obory. Polysémní termíny budou mít obecně vyšší relativní disponibilitu, protože jsou frekventovanější a zároveň se vyskytují ve více oborech, a tudíž i ve větším počtu textů.

[32] Vysoká hodnota ARF zároveň znamená, že dané slovo má i vysokou frekvenci.

[33] Polysémií a polyfunkčností se zabývají kapitoly 1.2.1.2, 1.2.2.1, 1.3.1.4 a 1.3.1.5.

mená, že pro jejich rozpoznávání můžeme využít tytéž **lexikální asociační míry**, které se běžně používají při hledání kolokací v korpusech. Tyto míry ukazují na kombinace dvou nebo více jednotek v korpusu, které se spolu vyskytují častěji, než by bylo dáno pouhou pravděpodobností (takové jednotky pak obvykle tvoří kolokaci). Hodnoty nejrůznějších asociační měr jsou obvykle spíše vodítkem při určování kolokací, pro větší spolehlivost je možné využít i kombinaci několika měr.

Vlastnosti vycházející z kolokačního chování víceslovných termínů jsou asociační míry MI-score a t-score (se značkami MWT:MI-score, MWT:t-score, viz kap. 2.2.4).

Kontextové vlastnosti

Kromě asociačních měr je možné k hledání víceslovných termínů (jako specifického typu kolokací) využít i kontextového chování slov, jako je proximita nebo obligatornost kontextu. Tyto vlastnosti zachycují, jak jsou slova v bigramu od sebe v textech obvykle vzdálená (proximita) a jak hodně si jedno z nich vyžaduje ve svém kontextu přítomnost druhého (obligatornost) (Cvrček, 2013, s. 36 a 43).

Mezi kvantitativní vlastnosti patří i zachycení **variability kontextu**, ať už vertikální (množství slov, která se můžou vyskytovat na určité pozici vůči zkoumanému slovu, Cvrček, 2013), tak horizontální (obvyklost n-tice slov sousedících se zkoumaným slovem, tamtéž). Ukazatelem variability kontextu je například entropie (míra neuspořádanosti zkoumaného systému, v tomto případě kontextu, Cvrček, 2013, s. 119) a v případě horizontální variability vážený průměr relativních frekvencí slov vyskytujících se v kontextu (tamtéž).

Konkrétní rysy vycházející z kontextového chování víceslovných termínů jsou Herfindahl-Hirschmanův index, entropie, vážený průměr relativních frekvencí předcházejícího kontextu, asociační míra obligatornost, asociační míra proximita, modus výskytu a absolutní frekvence kolokátu (se značkami KontextHH, KontextE, MWT:Oblig, MWT:Prox, MWT:Modus, MWT:AFC5 a MWT:AFC3, viz kap. 2.2.4).

Délka slova

Tato vlastnost je ve výzkumu použita (pod značkou Len_{syl}, viz kap. 2.2.4) na základě předpokladu, že mezi termíny se často vyskytují neobvykle dlouhá slova (*echokardiografický* nebo *glutamátoxalacetáttransamináza*). Délka slova (počítaná ve slabikách) tedy může ve vyhledávání termínů hrát určitou roli.

1.2.2.2 Kvalitativní vlastnosti termínů

Slovní druh

Příslušnost ke slovnímu druhu může být velmi dobře využita při automatickém vyhledávání termínů, a také často využívána je. Naprostá většina termínů patří totiž k podstatným jménům, případně i k přídavným jménům či slovesům. Víceslovné termíny jsou obvykle jmenné fráze (Kageura, Umino, 1996, s. 15), složené z podstatného a přídavného jména, a jsou tedy funkčně také substantivy (viz kap. 4.4).

Zastoupení ostatních slovních druhů v terminologii je pouze minimální, na druhou stranu v podstatě žádný slovní druh nelze z terminologie předem vyloučit, na což poukazuje i Hausenblas (1962) s tím, že např. některé předložky nebo spojky mají v některých případech specificky vymezený význam (*pět na druhou*, *jestliže..., pak...* apod.). Dokonce i některé částice mohou mít specifický význam v určité disciplíně (částice *Nechť* v matematice). Snad jen citoslovce jsou výjimkou a jako termíny si je lze představit jen těžko.

V korpusovém výzkumu je nutné se rozhodnout, zda zařazení do slovního druhu využít. Výhodou je, že korpusy v dnešní době obsahují poměrně velmi spolehlivé značkování (vč. zařazení do slovního druhu). Naproti tomu je nevýhodné, že nelze pracovat s neoznačkovanými texty. Pokud bychom tedy chtěli úspěšnou metodu ATR založenou zčásti na slovních druzích použít pro vyhledávání termínů v dalších textech, musely by tyto texty nejprve projít procesem značkování.

Neobvyklost formy (mezinárodnost)[34]

Chung (2003) zahrnuje do výčtu kvalitativních vlastností termínů (v angličtině) i latinský či řecký původ. Stejně tak v češtině existuje mnoho termínů cizího původu, nejčastěji přejatých z řečtiny a latiny (do češtiny obvykle přes jiný jazyk), v současné době také z angličtiny. Přejatá slova se často vyznačují formou neobvyklou pro češtinu. Tato neobvyklost (např. některá písmena v češtině málo používaná, neobvyklé kombinace písmen) může být vyjádřena kvantitativně, lze ji tedy použít při automatickém zpracování (v kap. 2.2.4 pod značkou Struct).

[34]Označení mezinárodnost používá Poštolková et al. (1983, s. 67-73).

Jednoznačnost

Jedním z častých požadavků na termín je, aby byl jednoznačný, tedy nikoli polysémní[35] (Čermák, 2010, s. 132; Poštolková et al., 1983, s. 79). Některé termíny nabývají v jiném typu textů, v jiném oboru, ale někdy dokonce v tomtéž oboru dalších významů, nebo se jeho význam nějakým způsobem mění[36].

Polysémie je do značné míry schopná ovlivnit výsledky automatického vyhledávání termínů. Polysémní slova se frekvenčně a distribučně chovají odlišně od monosémních[37]: budou mít s vysokou pravděpodobností vyšší frekvenci v obecném jazyce a/nebo vyšší distribuci v rámci zkoumaných oborů (Bečka, 1972, s. 58).

Definovanost

Požadavek na definovanost termínu je pochopitelný a v některých případech i oprávněný (např. u preskriptivních terminologií, viz kap. 1.2.1.1) z pohledu terminologické standardizace, ale ne vždy z pohledu automatického vyhledávání termínů. To je totiž zaměřeno nikoli jen na termíny ustálené, takové, které už můžeme najít v terminologických slovnících, ale právě také na termíny nové, ještě ne zcela zavedené, neustálené.

Definovanost (viz kap. 1.2.1) jako rys velké části termínů však může být využita při přípravě materiálu pro automatické vyhledávání termínů. Při ručním označování termínů a netermínů v trénovacích datech pro tuto práci se využívá terminologických slovníků jednotlivých zkoumaných oborů.

Ustálenost

Dalším rysem připisovaným termínům je jejich ustálenost (Čermák, 2010, s. 132; Poštolková et al., 1983, s. 62). Za ustálené lze pokládat takové termíny, které jsou pevnou součástí systému termínů v daném oboru. Ustálenost je nabývána územ, případně i definováním

[35] V praxi není tato podmínka vždy splněna, což může pro automatické vyhledávání termínů představovat nepříjemný problém – úspěšnost dané metody se tím snižuje, zvlášť vyskytne-li se takový případ v trénovacích datech, na kterých velmi záleží. Lze uvést příklad slova *špičkový* v energetice: dokonce v jediném textu se objevuje *špičkový* jako součást víceslovného termínu *špičkový výkon* i jako součást běžné kolokace *špičkový odborník*.

[36] Velká většina termínů sice polysémní není (jde o velký počet vysoce specializovaných termínů), ale z lingvistického (zvláště pak lexikografického) hlediska jsou zajímavé právě ty jednotky, které polysémní jsou (mj. proto, že bývají poměrně frekventované).

[37] Příkladem může být lexém *strom*: v neformálním rozhovoru většinou nebude mít terminologickou platnost, ale termínem bude v botanice, matematice (typ grafu) a dalších oborech, jako součást víceslovných termínů se navíc objeví i v sociologii (*strom významnosti*), biologii (*molekulární strom*), lingvistice (*generativní strom*) atd. V tomto případě může bohatá polysémie velmi ztížit automatické vyhledání termínu.

termínu nebo jeho zahrnutím do závazné normy oboru, v některých oborech tím, že je termín součástí nomenklatury.

Některé termíny jsou přijímané více méně všeobecně, jiné se používají jen ad hoc v rámci jednoho textu (Bečka, 1972, s. 48), ještě další mají mnoho vzájemně si neodpovídajících definic - k tomu často tíhnou zejména termíny ve vědách humanitních (Teubert, Čermáková, 2004; Machová, 1995; Kocourek 1965; k tomu viz i kap. 1.2.1.1. Nejvyšší míru ustálenosti zpravidla vykazují názvoslovné termíny, které mají své pevné místo v závazném názvosloví (nomenklatuře) oboru[38], nebo termíny zahrnuté v některé závazné normě.

Ustálenost v jiném smyslu hraje roli při rozpoznávání víceslovných termínů, které jsou systémovými kolokacemi - to znamená, že jejich víceslovná forma je ustálená. Při vyhledávání víceslovných termínů je možno použít nejrůznější lexikální asociační míry, které jsou schopny určit, jak těsné (ustálené) je spojení jejich jednotlivých členů (viz kap. 1.2.2.1).

1.3 Princip škály v terminologii

Při zkoumání terminologie je velmi důležité si uvědomit, že termíny mají škálovitou povahu (Čermák, 2010, s. 134): terminologická platnost některých slov je velmi silná, u jiných zase slabá, a existují i slova, která mají terminologickou platnost nulovou (do této skupiny bude patřit většina gramatických slov).

Jedním z významných faktorů ovlivňujících sílu terminologické platnosti je typ textu, v kterém se dané slovo vyskytuje (např. akademický vs. publicistický text), případně i konkrétní obor. Primárním místem výskytu termínů jsou akademické texty (resp. texty odborné, tzn. i texty oborů praktických). Termíny se ale ve skutečnosti můžou vyskytovat a vyskytují v jakémkoli typu textu, odborném či neodborném, dokonce i v běžné každodenní komunikaci, ve spontánních rozhovorech: Těžko můžeme například o slovech jako *antibiotika* nebo *software*, která se vyskytují i v běžném hovoru, říct, že se nejedná o termíny (o termínech v neodborných textech viz Ville-Ometz et al., 2007, s. 37). S textovými typy souvisí i otázka terminologizace běžných slov a determinologizace oborových termínů - slova pronikají z jedné sféry do druhé a jejich terminologická platnost se buď posiluje, nebo oslabuje (o de/terminologizaci viz Čermák, 2010, s. 134).

Terminologická platnost slov se při výzkumu obvykle pohybuje na škále, kde 0 je nulová

[38] Výjimkou jsou termíny, které jsou velmi řídké, ale díky své systémovosti jsou do terminologie oboru zahrnuty. Zde jdou hlediska ustálenosti a systémovosti proti sobě.

terminologická platnost a 1 je nejsilnější (příp. absolutní) terminologická platnost[39]. Pozici na škále ovlivňují jak vlivy vnější, nezávislé na chování termínů v textech, tak samotné vlastnosti jednotlivých termínů.

Vnější vlivy ovlivňující sílu terminologické platnosti:

- **odbornost textu**, ve kterém se vyskytují (kap. 1.3.1.1)
- přináležitost termínu do oboru, do kterého text patří: **oborová příslušnost** (kap. 1.3.1.2)

Vlastnosti termínů ovlivňující sílu terminologické platnosti:

- **definovanost termínu** (kap. 1.3.1.3)
- distribuční a frekvenční chování, především jejich **průnik** do obecných textů (**polysémie**) a příslušnost k více oborům v různých významech (**polyfunkčnost**) (kap. 1.3.1.4 a 1.3.1.5)
- u jednoslovných termínů **neobvyklost formy** (vč. délky slova) (kap. 1.3.1.6)
- u víceslovných termínů[40] fakt, zda se jedná o **kolokaci** a většinou i **přítomnost alespoň jednoho termínu** v kolokaci

Výše zmíněné charakteristiky mají nějakým způsobem vliv na automatické zpracování termínů, některé přímo, a jiné nepřímo. Nepřímý vliv má odbornost textu, a to při výběru textů pro výzkum. Při automatickém vyhledávání se obvykle pracuje s odbornými texty, jako jsou abstrakty nebo databáze vědeckých článků (např. Nobata et al., 1999) či vysokoškolské učebnice (Chung, 2003), a srovnává se např. výskyt jednotlivých slov v odborných a beletristických nebo publicistických textech (Yang, 1986; Šrajerová, 2009a,b; Šrajerová et al., 2009). Nepřímo ovlivňuje vyhledávání termínů také definovanost termínů a zároveň i jejich oborová příslušnost, a to při přípravě dat k automatickému zpracování -

[39] Je nutné rozlišovat škálu založenou na intuitivním vnímání termínu, která není zcela spolehlivá (k tomu viz i Cvrček a Kováříková (2011, s. 117) o malé spolehlivosti lingvistického zkoumání na základě introspekce), a škálu reálnou, založenou na výzkumu reálných dat, například na výsledcích automatického vyhledávání termínů, která je zase daleko méně přehledná; mj. nelze na jejím základě jednoznačně a obecně určit, co je a co není termín, je nutno si stanovit vlastní hranici, která odpovídá cílům konkrétního výzkumu (na škále 0 až 1 to může např. být hranice na hodnotě 0,5, více v kap. 3.3.1 nebo 3.3.2).

[40] Víceslovné termíny mají menší tendenci pronikat do obecných textů a obvykle bývají spojeny s jedním oborem (příp. několika příbuznými). I neobvyklost formy se posuzuje jen u jednotlivých složek víceslovných termínů.

v této práci se termíny v trénovacích datech vybírají na základě terminologických slovníků, kde jsou zaznamenány v oboru ustálené termíny většinou i s definicemi.

Přímo se na automatickém vyhledávání podílejí frekvenční a distribuční vlastnosti daného slova nebo kolokace, neboť právě na jejich základě se v rámci automatického zpracování rozhoduje, zda se jedná, či nejedná o termín (resp. jak silná je jeho terminologická platnost). Přímý vliv můžou mít i rysy jako neobvyklost formy či počet slabik, u víceslovných termínů pak přítomnost alespoň jednoho termínu v kolokaci a zároveň hodnota asociační míry.

1.3.1 Charakteristiky ovlivňující sílu terminologické platnosti

1.3.1.1 Odbornost textu

Termíny se primárně vyskytují v oficiálních odborných textech psaných i mluvených, jako jsou vědecké články či monografie, učebnice, vysokoškolské přednášky i jiná odborná výuka, prezentace apod., a také v databázích či normách z těchto textů odvozených. Bečka (1972, s. 48) poukazuje na to, že přesné užívání slov s plnou terminologickou platností (stejně jako porozumění takovým slovům) předpokládá vědeckou znalost jevu v daném vědeckém oboru. Počítá přitom s tím, že termíny s plnou terminologickou platností se vyskytují pouze v akademických textech. Čermák (2010, s. 132) upozorňuje na výskyt silných termínů nejen v rámci vědeckých oborů, ale i v oborech praktických či v řemeslech. Bečkovo tvrzení by tedy bylo vhodné upravit tak, že přesné a plnohodnotné používání a porozumění termínům v odborných textech vyžaduje specializovanou znalost v daném oboru, ať už vědeckém, nebo praktickém.

Taková specializovaná znalost však stále ještě nezaručuje přesné používání termínů (tedy s plnou terminologickou platností), je jen jeho předpokladem[41]. Oficiálnost textu je dalším takovým předpokladem, protože právě v oficiálních textech je kladen důraz na přesnost vyjadřování, které umožňuje i přesnost pochopení daného textu.

Termíny se používají nejen v oficiálních odborných textech, ale v podstatě v jakémkoli textu[42] včetně neformálních mluvených dialogických textů[43]; v kterémkoli textu jiném než

[41] Je možné si například představit poměrně běžnou situaci, kdy dva odborníci znají sice přesný význam některého termínu, v neformální situaci ho ale z úsporných důvodů používají nepřesně, např. ve zúženém významu.

[42] K tomu i Encyklopedický slovník češtiny: „Termíny se uplatňují nejen ve sféře odborného vyjadřování, ale také ve sféře publicistické a běžného dorozumívání.“ (Karlík et al., 2002, s. 488, heslo Termín)

[43] To je typ textů, jaké zaznamenávají mluvené korpusy ČNK: ORAL2006, ORAL2008 a ORAL2013.

oficiálním odborném se ale snižuje síla terminologické platnosti daného termínu, protože se snižuje přesnost jeho použití i pochopení a jeho význam se zužuje, rozšiřuje nebo posunuje[44] (viz kap. 1.3.1.4).

1.3.1.2 Příslušnost termínu k oboru

Síla terminologické platnosti termínu nezávisí jen na odbornosti textu; velkou měrou se na ní podílí také přináležitost termínu do určitého oboru. Za termín se silnou terminologickou platností lze pokládat takový termín, který patří do systému termínů v oboru, jehož součástí je konkrétní zkoumaný text. Čím je obor, kam daný termín patří, vzdálenější[45] od oboru, k němuž náleží zkoumaný text, tím je terminologická platnost termínu slabší. Pokud tedy použijeme v botanickém textu termín z oblasti chemie, jeho terminologická platnost je oslabena oproti termínům, které náležejí do oblasti botaniky. Pokud v tomtéž textu použijeme termín z oblasti dějin umění, jeho terminologická platnost bude velmi slabá, protože tyto dva obory jsou velice vzdálené[46].

Příslušností termínu do konkrétního oboru se zabývali Kageura a Umino, kteří ho označili jako termhood (termínovost). Termínovost podle něj závisí na tom, jak těsně lingvistická jednotka souvisí s konceptem v dané disciplíně (Kageura, Umino, 1996, s. 11). Podobně Chung zmiňuje mezi kritérii pro určování termínů (v rámci automatického vyhledávání) vztah významu termínu ke konkrétnímu specializovanému oboru (Chung, 2003, s. 221). Stejně tak jsou si důležitosti spjatosti termínu s daným oborem intuitivně vědomi i odborníci (nelingvisté) z různých disciplín, mají-li termíny v textu vyhledat ručně (Machová, 1995, s. 139).

Oborově nespecifické vědecké termíny

Zvláštním případem jsou vybrané termíny z oblasti filozofie - teorie vědy, statistiky apod., které se staly součástí obecnější akademické slovní zásoby. Jsou to slova jako *analyzovat, přístup, funkce, systém, teoretický*, ale i kolokace jako *kvalitativní výzkum, analýza dat,*

[44] Příkladem významového posunu může být slovo *ekolog*, což je v biologii termín pro odborníka zabývajícího se vztahem organismů a jejich prostředí a vztahem organismů navzájem, kdežto v běžném slovníku je obvykle používáno pro ochranáře životního prostředí (viz i kolokace slova *ekolog* v SYN2010: *ochránce, záchrana, protestovat, radikální, aktivista* apod.).

[45] Problémem zůstává, jak určit příbuznost či vzdálenost jednotlivých oborů (jinak než intuitivně). Jak se ukazuje, právě automatické vyhledávání termínů a aplikace jeho výsledků takovouto informaci do určité míry poskytuje (Šrajerová, 2009b). Zde viz kap. 4.2.

[46] V případě textu, který náleží do více oborů (je interdisciplinární), se pak zkoumá přináležitost termínu aspoň k jednomu z těchto oborů.

tvrdá data, nulová hypotéza apod. Patří mezi termíny se slabou terminologickou platností, které mají v rámci akademických textů vysokou frekvenci i distribuci (jednoduše řečeno: vyskytují se často a ve všech oborech), ale obvykle bez tématické příslušnosti k danému oboru.

Z takovýchto slov je složen tzv. **Academic Word List**[47] (AWL) (Coxhead, 2000), který byl vytvořen pro potřeby vysokoškolských studentů, především těch, jejichž mateřským jazykem není angličtina. Seznam obsahuje 570 hesel (jejich součástí jsou slova příbuzná s heslem), a ta jsou rozdělena do 10 skupin podle frekvence. Jednotlivé složky AWL pokrývají podle Coxhead (2000, s. 226) více než 10 % korpusu akademických textů sestaveného pro daný výzkum. Seznam AWL je někdy kritizován pro přílišné zobecňování, protože akademické obory jsou velice odlišné a každý má svoje specifika (Hyland, Tse, 2007).

Touto skupinou slov se zabývá také Da Sylva, která ji nazývá **Basic Scientific Vocabulary**[48] (BSV) (Da Sylva, 2009). Součástí tohoto seznamu je zhruba 4000 slov. I tato autorka kritizuje AWL, tentokrát na základě zvolené metody: z AWL je vyřazeno každé slovo, které patří do seznamu 2000 nejfrekventovanějších slov (či spíše hesel zahrnujících i příbuzná slova) angličtiny (General Service List, West, 1953), tedy i slova jako *schopnost*[49] nebo *zkoumat* či *studovat/studie*[50] (Da Sylva, 2009, s. 11).

Seznamy, jako jsou AWL nebo BSV, zachycují jeden z největších problémů vyhledávání termínů (viz i kap. 2.2.5.2). Jedním z nejsložitějších úkolů při ručním vyhledávání termínů je totiž rozlišit, kdy jde o termín se silnou terminologickou platností (náležející do daného oboru), a kdy se jedná o oborově nespecifické termíny s nižší terminologickou platností. Lze předpokládat, že takové termíny by bylo možné vyhledávat automaticky podobnou metodou, jaká je prezentována v této práci, ovšem na základě jiných kritérií, nebo spíše jiných hodnot distribučních a frekvenčních charakteristik. Pokud by se to podařilo, vznikl by základ pro specializovanou (vědeckou) část hesláře obecného slovníku[51].

[47] Seznam akademických slov, překlad autorky.

[48] Základní vědecká slovní zásoba, překlad autorky. Da Sylva rozlišuje čtyři druhy slov v akademických textech: vedle tzv. prázdných (gramatických) slov jsou to tři třídy plnovýznamových slov - common vocabulary (běžná slova), basic scientific vocabulary (základní vědecká slovní zásoba) a specialized scientific and technical vocabulary (specializovaná vědecká a technická slovní zásoba) (2009, s. 2-3). Naproti tomu např. Bečka (1972, s. 47) vedle gramatických slov rozeznává pouze slova terminologická a neterminologická.

[49] Angl. orig.: *ability*.

[50] Angl. orig.: *study*.

[51] Součástí specializované části hesláře by pak nutně musely být ještě „tématické“ termíny z jednotlivých oborů, vybrané na základě své frekvence a distribuce v textech odborných a obecných.

1.3.1.3 Definovanost/ustálenost termínu

Třetím kritériem, které ovlivňuje škálovitost termínu, je definovanost či ustálenost termínů. Definovanost a ustálenost nepovažuji za nutnou podmínku termínu (viz kap. 1.2.1.1 a 1.3.1.3). Tím, že je termín definován a/nebo ustálen (a je např. uveden v nějakém terminologickém slovníku nebo v závazné normě), se však zvyšuje jeho terminologická platnost v daném oboru. Definovanost nebo ustálenost termínu není možné kvantifikovat, proto jich nelze využít při samotném procesu automatickém zpracovávání terminologie. Nicméně obě vlastnosti mají význam při přípravě materiálu pro ATR (viz kap. 2.2.5.1).

1.3.1.4 Průnik do obecných (neodborných) textů

Terminologická platnost termínu se zvyšuje nebo snižuje v závislosti na tom, v jakém typu textu se termín vyskytuje. Termíny, které lze najít jak v odborných textech, tak v textech obecných (ať už v publicistice, v beletrii nebo třeba v neformálních mluvených projevech) nejsou ojedinělé, ale v rámci milionů odborných termínů[52] jde vlastně o okrajovou záležitost[53]. Oproti vysoce specializovaným termínům vyskytujícím se především (nebo dokonce výhradně) v odborných textech se budou chovat jinak zvláště z hlediska rozložení v korpusu a z hlediska frekvence (poměr jejich frekvence v odborných textech ku frekvenci v textech obecných bude výrazně nižší, viz kap. 5.1.1). Odlišovat se bude i jejich rozložení v korpusu (vyjádřené např. jako ARF).

U termínů, které pronikají do obecných textů, jde o zvláštní druh polysémie (Čermák, 2010, s. 134), kdy se význam termínu zužuje, rozšiřuje, či posunuje oproti užití v textech odborných. V souvislosti s výskyty v neodborných textech se mluví o **determinologizaci**[54] (Čermák, 2010, s. 134; Karlík et al., 2002, s. 488, heslo Termín): některé termíny se začnou užívat i mimo původní obor, nejprve v příbuzných oborech, v části případů pak i v popularizující literatuře a publicistice - a jejich terminologická platnost tím oslabuje. Proces může pokračovat dál, původní termín se může začít objevovat v obecných textech (např. v beletrii nebo v neformálním rozhovoru), kde je jeho význam už obvykle značně odlišný

[52] K počtu termínů viz Čermák (2010, s. 135).

[53] Pokud bychom ale sestavovali heslář pro slovník češtiny (viz kap. 1.1.1), právě takovéto termíny nás budou zajímat, protože jsou díky své přítomnosti v obecných textech známější a používanější, a tudíž je vyšší pravděpodobnost, že je mluvčí češtiny budou ve slovníku vyhledávat.

[54] O determinologizaci lze uvažovat jako o procesu, nebo jako o výsledku procesu (Čermák, 2010, s. 134). Oba případy by bylo možné rozlišit i terminologicky: determinologizace jako proces a determinologizovanost jako výsledek tohoto procesu.

od původního terminologického (např. *stres* či *deprese*), a jeho terminologická platnost je velmi slabá. Je otázkou interpretace, zda v některých případech dochází k úplné determinologizaci termínu, kdy by se původní termín stal běžným slovem s nulovou terminologickou platností.

Opačný proces, **terminologizace** běžného slova, není takto postupný: některá disciplína, nebo např. některý praktický obor, převezme slovo z obecného jazyka a používá ho (obv. v zúženém významu) jako termín. To se týká většiny konkrétních substantiv, vždy totiž existuje alespoň jedno odvětví lidské činnosti, které dané substantivum používá jako termín (např. *strom, voda, sůl*) (Čermák, 2010, s. 134; Karlík et al., 2002, s. 488, heslo Termín).

1.3.1.5 Polyfunkčnost (mezioborová polysémie)

Určitá část termínů se vyskytuje ve více oborech, a to buď s významem víceméně stejným jako výpůjčka z oboru příbuzného, nebo s významem odlišným - ve druhém případě jde o polyfunkčnost či mezioborovou polysémii[55]. Příkladem mezioborové polysémie může být termín *komunikace*, který má jiný význam v lingvistice, v botanice a v dopravě (viz kap. 1.2.1.2 a 1.2.2.2). Mezioborová polysémie se projeví v distribučním chování termínu - daný termín se objevuje ve větším množství oborů, a to i nepříbuzných (např. lingvistika a doprava).

1.3.1.6 Neobvyklost formy

Dalším aspektem, který může mít výrazný vliv na sílu terminologické platnosti, je neobvyklost formy termínu, a to jak z hlediska pro češtinu nezvyklých písmen nebo kombinací písmen, tak z hlediska neobvyklé délky slova (Šrajerová, 2009a, s. 14); viz i kap. 1.2.2.1 a 1.2.2.2. Příkladem může být lingvistický termín *monoftongizace diftongů*.

1.3.2 Škála v terminologii a prototypický termín

Prototypický termín (z hlediska škály termín s vysokou terminologickou platností) se tedy vyskytuje v odborných (nejspíše akademických) textech jediného oboru, do jehož terminologie přináleží. Je jednoznačně definovaný nebo v oboru jinak ustálený (vyskytuje se např.

[55] Ani v prvním případě ale nelze vyloučit jistý stupeň polysémie, neboť význam může být pozměněný oproti užití v „domácím“ oboru.

v terminologickém slovníku nebo v závazné normě daného oboru). Má neobvyklou formu ve srovnání se slovy užívanými v neodborných textech. V případě víceslovného termínu jde o kolokaci.

2 Materiál a metoda

Tato kapitola se soustředí na obecné principy data miningu, který je základem metody automatického vyhledávání termínů TERMIT, a na vysvětlení vybrané terminologie (kap. 2.1). Druhá část kapitoly (kap. 2.2) je věnovaná podrobnému popisu práce s materiálem připravovaným pro data mining: výběru a zpracování korpusového jazykového materiálu, vlastnostem, které jsou přidělovány jednotlivým textovým pozicím[1] (obv. slovům) v trénovacích i testovacích datech, ať už ručně, nebo automaticky (nejrůznější statistické, kontextové a lingvistické rysy), ručnímu označování termínů v trénovacích datech a nejvhodnější podobě dat (lemma vs slovní tvar apod.). Třetí část (kap. 2.3) se soustředí na samotnou metodu TERMIT a popisuje jednotlivé kroky, které vedou k identifikaci termínů v reálných textech. Na závěr jsou představeny způsoby vyhodnocování úspěšnosti metody TERMIT při vyhledávání termínů.

2.1 Data mining

Data mining je charakterizován jako proces hledání opakujících se vzorů ve velkých objemech dat za pomoci počítačových algoritmů (Witten, Frank, 2005, s. 5). Jde o data takového rozsahu, že není možné je zpracovat jinak než automaticky (v našem případě jde o tisíce textových pozic v korpusu, nejčastěji slov, přičemž každá pozice má přiřazeny desítky vlastností). Základním požadavkem na taková schémata objevená v datech je to, aby byla nějakým způsobem užitečná, tedy aby vysvětlovala nějaký jev nebo přinesla nový smysluplný poznatek o datech. Smysluplné vzory nám umožňují „vytvářet netriviální předpovědi pro nová data“ (Witten, Frank, 2005, s. 5) - na základě data miningu je tedy např. možné s určitou pravděpodobností automaticky předpovídat, které textové pozice v předložených textech jsou termíny.

V data miningu dochází nejprve k automatickému učení, což je proces, při kterém algoritmus hledá pravidelnosti či schémata. Zde užívaný učící proces se nazývá „učení s učitelem“[2] - v jeho rámci je předem stanoven výsledek. V případě předkládané studie je v trénovacích

[1] V korpusové terminologii je textová pozice znak nebo řetěz znaků oddělený z obou stran mezerou, nejčastěji slovo. Interpunkční značky jsou obvykle samostatnou textovou pozicí.

[2] Ve druhé kategorii učících metod, tzv. učení bez učitele, jsou data-miningovému nástroji poskytnuta data; jeho úkolem je najít v těchto datech jakékoli předem neurčené, ale užitečné vzory, např. rozdělit je do soudržných skupin s co nejodlišnějšími vlastnostmi (v lingvistice by se např. dalo uvažovat o automatickém rozdělování do slovních druhů, které by ovšem ve výsledku mohly být zásadně odlišné od současných slovních druhů).

datech předem dáno, které z textových pozic jsou termíny. Úkolem data miningu je zjistit, jakým způsobem se k tomuto výsledku dobrat. Na základě toho je pak v druhé fázi možné předpovědět výsledky na dalších (testovacích) datech, tj. automaticky vyhledat termíny (s určitou úspěšností) v dalších poskytnutých textech.

Badatel z jiného oboru, než je výpočetní technika, nemusí nutně rozumět všem podrobnostem fungování data-miningového nástroje. Mnohem důležitější je porozumění tomu, co jsou vstupní a výstupní data a jak s nimi pracovat. Velmi zjednodušeně se dá říct, že badatel připraví materiál (vstup) a data-miningové nástroje mu poskytnou výsledky (výstup). Následuje analýza a interpretace výsledků z hlediska dané disciplíny (lingvistiky, terminologie). Příprava vstupních dat je náročným procesem (ten je podrobně popsaný v kap. 2.2), na jejich kvalitě totiž závisí kvalita výsledků použitých metod.

Některé termíny z oblasti data miningu, které jsou zde používány:

vstup/vstupní data: data ve formě rozsáhlé tabulky (matice, viz níže) poskytnutá data-miningovému nástroji pro učení;

výstup/výstupní data: výsledky procesu data miningu: nalezená pravidla, pomocí nichž je možné přiřadit stupeň/hodnotu terminologické platnosti k libovolné textové pozici; výstupem může být i ohodnocení důležitosti jednotlivých rysů pro proces učení nebo ohodnocení úspěšnosti použité metody;

instance: instanci zde odpovídá textová pozice[3] v korpusu (buď ve formě slovního tvaru, nebo lemmatu). U víceslovných termínů je instancí bigram (dvě po sobě následující textové pozice);

atribut: statistická nebo lingvistická vlastnost přiřazená ručně nebo automaticky instancím/textovým pozicím (např. počet slabik);

hodnota atributu: hodnota, kterou daný atribut vykazuje u jednotlivé instance/textové pozice (např. 3 slabiky);

matice: rozsáhlá tabulka, kde v řádcích jsou jednotlivé instance/textové pozice a ve sloupcích jsou instancím přiřazovány hodnoty jednotlivých atributů (vlastností);

[3]O jednotkách v korpusu se v souvislosti s procesem tokenizace nemluví jako o slovech, ale jako o (textových) pozicích. Textová pozice je řetězec znaků (grafické slovo) oddělený z obou stran mezerou. Výjimku tvoří interpunkční znaménka, která jsou také samostatnými (textovými) pozicemi.

trénovací data: vstupní data, na nichž se data-miningový nástroj naučí a vytvoří pravidla pro stanovení terminologické platnosti (příp. ohodnotí důležitost jednotlivých atributů);

testovací data: vstupní data, na nichž se v závěru testuje, jak úspěšně je nástroj schopen vyhledávat termíny v nových datech.

2.2 Materiál

Příprava materiálu (vstupních dat) pro automatické vyhledávání termínů je velmi složitým procesem. Pro vytvoření kvalitních a relevantních vstupních dat je třeba podniknout následující kroky:

- výběr vhodného korpusu
- vytvoření srovnávacího subkorpusu neakademických textů (beletrie a publicistika)
- vytvoření subkorpusu odborných textů a subkorpusů jednotlivých disciplín
- výběr trénovacích a testovacích dat pro data-miningové nástroje (z textů vybraných oborů)
- ruční označení termínů a netermínů v trénovacích datech
- automatické přiřazení hodnot dalších rysů (lingvistických, statistických, kontextových) trénovacím datům a testovacím datům - vytvoření matice.

Velká pozornost je v této kapitole věnována především ručnímu označování termínů (které se ukázalo být nejproblematičtější částí výzkumu) a dále všem rysům přiřazovaným textovým pozicím v trénovacích datech.

2.2.1 Výběr vhodného korpusu

Z korpusů ČNK byl pro tento výzkum vybrán jako nejvhodnější referenční korpus SYN2010, reprezentativní synchronní korpus psaných textů. Výhodami tohoto korpusu jsou snadná dostupnost dat, reprezentativnost (jsou k dispozici texty odborné i beletrie a publicistika, a to ve víceméně vyváženém poměru) a ideální velikost: množství dat je dostatečné, ale není

přílišné - ve větším korpusu, např. SYN, by bylo třeba texty použitelné pro daný výzkum daleko víc třídit. Kromě toho je přínosem i možnost zpracování některých atributů pro trénovací data (konkrétně jde o ARF, průměrnou redukovanou četnost, viz níže), které by ve větším a nevyváženém korpusu bylo složitější a méně vypovídající.

Obecně je k lingvistickým výzkumům výhodné využívat korpusy referenční (jako je právě SYN2010), tj. takové, které se po svém zveřejnění už nemění, protože experimenty je pak možné vždy zopakovat za stejných podmínek, tj. se zcela identickými daty[4].

Jednou z předností korpusů ČNK je jejich anotace, pomocí níž jsou označeny jednotlivé dokumenty a jejich zařazení mezi textové typy, stejně jako jejich žánrová příslušnost. Díky tomuto zpracování dat je možné vybrat například jen texty publicistické či jen odborné[5], nebo texty z určitého oboru - to je pro výzkum terminologie zcela zásadní. V neposlední řadě může být užitečná i anotace lingvistická, tj. lemmatizace a morfologické značkování. Při výzkumu je možné, a někdy i vhodné, tohoto značení nevyužít, nicméně vždy je výhodné mít k dispozici co nejvíc informací, z nichž pak můžeme vybírat.

Akademické texty jsou v korpusech ČNK označeny jako vědeckonaučné a najdeme je pod textovým typem označeným zkratkou SCI. V korpusu SYN2010 jsou k dispozici akademické texty z celkem 37 oborů[6] - ty lze mezi sebou na základě výsledků výzkumu porovnávat (množství termínů, typy termínů apod., viz kap. 4.1) nebo mezi nimi lze zjišťovat vzájemné vztahy. Seznam všech použitých oborů je v tabulce 2.1.

Nevýhodou využití korpusových dat jsou chyby, ke kterým při anotaci v určité míře vždy dochází. Při výzkumu terminologie představuje velký problém text omylem zařazený do nesprávného textového typu, případně do jiného oboru[7]. I nesprávná lemmatizace nebo morfologické značkování, které jsou jinak velmi dobrým nástrojem, můžou mít vliv na

[4]To by částečně bylo možné zajistit i u korpusu nereferenčního přesným popisem použitých dat (výpisem veškerých textů, které byly v daném výzkumu použity). Měnící se anotace však může výsledky přesto zkreslit.

[5]Ne všechny odborné texty v SYN2010 jsou čistě akademické, řadí se sem i učebnice, encyklopedie, populárněnaučná literatura, zájmové časopisy a administrativní texty.

[6]Ve skutečnosti je oborů obsahujících akademické texty celkem 44, ale 7 z nich bylo z dat z různých důvodů vyřazeno (např. LAW se specifikací oboru „jiný z oblasti práva a bezpečnosti“).

[7]Při vybírání trénovacích dat se ukázalo, že správná anotace textů v korpusu, tj. jejich zařazení do textového typu a do určitého oboru, velmi ovlivňuje průběh úzce zaměřených výzkumů, jako je výzkum terminologie. Z 35 textů (všechny patří do korpusu SYN2010) ve zkoumaných čtyřech oborech je 17 zařazeno do nesprávného textového typu (většinou se jedná o učebnice/skripta) nebo do nesprávného oboru. Je ale nutno podotknout, že zařazování do textového typu nebo do oboru je poměrně náročné, protože se vnímání odbornosti textu mění obor od oboru (např. eseje jsou v literární vědě přijatelnou formou odborného textu, kdežto v technických oborech bychom eseje řadili spíš do popularizační literatury).

Tabulka 2.1: Seznam oborů v korpusu SYN2010, jejichž texty jsou zařazeny do textového typu vědeckonaučná literatura (SCI), a počet textových pozic, které tyto obory v daném textovém typu obsazují.

Zkratka	**Obor**	**Textové pozice**
AGR	zemědělství, lesnictví	7154
ANT	antropologie	33544
ART	výtvarné umění	429089
BIO	obecná biologie	206906
BOT	botanika	94848
CHE	chemie	140313
CIN	film	72603
COM	informatika	300021
ECO	ekonomie, obchod	396386
EDU	pedagogika	269774
ENE	energetika	103628
ENV	ekologie	370210
ETH	etnografie	121902
GEO	geologie	83582
HIS	historie	771311
IND	průmysl, technika	193058
INF	knihovnictví, informace	122605
JUR	právo	786944
LIN	lingvistika	154261
LIT	literární věda	297810
LOG	logika	84592
MAN	management, řízení	400122
MAT	matematika	132554
MED	lékařství	313502
MIL	vojenství	118265
MUS	hudba	142283
NAT	jiný z oblasti přírodních věd	233641
PHI	filozofie	452095
PHY	fyzika	143493
POL	politologie	210819
PSY	psychologie	268463
REL	náboženství, teologie	393693
SOC	sociologie	620621
SPO	sport	153795
THE	divadlo, balet	116272
TRA	doprava, telekomunikace	146296
ZOO	zoologie	122311
CELKEM		**9.008.766**

Tabulka 2.2: Seznam textových typů zahrnutých v subkorpusu COMPAR a počet textových pozic.

Zkratka	Typ textu	Textové pozice
NOV	román	29398037
COL	soubor povídek, jednotlivá povídka	8068491
VER	básně	1264519
IMA	jiné imaginativní texty	347012
PUB	publicistika	39401123
CELKEM		**78.479.182**

přesnost výsledků experimentů (k lemmatizaci a morf. značkování viz kap. 2.2.6).

2.2.2 Subkorpusy

Pro předkládaný výzkum bylo třeba vytvořit subkorpus akademických textů (textový typ SCI), který se skládá ze subkorpusů jednotlivých oborů a srovnávací subkorpus neakademických textů (beletristických a publicistických), nazývaný dále zjednodušeně srovnávací korpus. Pro první subkorpus je dále používáno označení **SCI** a pro srovnávací korpus označení **COMPAR**.

Akademický subkorpus SCI je využíván jednak pro výběr trénovacích a testovacích dat, jednak pro výpočty některých vlastností jednotlivých slov (např. distribuce - v kolika akademických oborech se dané slovo vyskytuje), a jednak posléze pro porovnávání jednotlivých oborů na základě vyhledaných termínů.

Srovnávací subkorpus COMPAR je nezbytný pro výpočty jednotlivých rysů, např. pro srovnání frekvence slova v akademických a neakademických textech. COMPAR se skládá ze všech textů zařazených v SYN 2010 do textových typů publicistických a beletristických (tab. 2.2).

2.2.3 Výběr trénovacích dat pro data mining

Pokud chceme při výzkumu používat data-miningové nástroje, musíme pro ně připravit trénovací data, tj. data, pomocí kterých se jednotlivé metody natrénují na vyhledávání termínů.

Na základě předchozích zkušeností byla trénovací data vybírána z menšího počtu oborů

Tabulka 2.3: Pro výzkum byly vybrány čtyři obory tak, aby byl materiál co nejrozmanitější. Byl vybrán jeden obor technický (COM), jeden přírodovědný (MED), jeden humanitní (LIT) a jeden sociálněvědný (SOC). Ve třetím sloupci je počet textů, které jsou v daném oboru k dispozici, a ve čtvrtém sloupci počet textových pozic, které tyto texty obsazují.

Zkratka	Obor	Texty	Textové pozice
COM	informatika	10	300021
LIT	literární věda	5	297810
MED	lékařství	9	313502
SOC	sociologie	7	620621

(čtyři obory), ale zato je každý z vybraných oborů zastoupen poměrně vysokým počtem textových pozic. Zároveň jsou textové pozice vybírány z většího počtu textů dané disciplíny. Trénovací data obsahují celkově osm tisíc textových pozic (vč. interpunkce), tj. dva tisíce pozic z každého oboru, přičemž data z každého oboru jsou složena z 500 po sobě jdoucích[8] pozic náhodně vybraných ze čtyř různých textů[9]. Celkově tedy data pocházejí ze 16 různých vědeckých monografií obsažených v korpusu SYN2010.

Trénovací data byla vybrána ze čtyř oborů (viz tab. 2.3), jednoho přírodovědného (MED - lékařství), jednoho technického (COM - informatika), jednoho humanitního (LIT - literární věda) a jednoho sociálněvědného (SOC - sociologie). Rozmanitost oborů je záměrná, lze totiž předpokládat, že terminologie takto různých oborů budou velmi odlišné a při hledání obecného vzorce pro automatické vyhledávání termínů v textech je třeba tuto odlišnost vzít v úvahu. Kromě rozmanitosti byly obory vybírány i na základě toho, jestli jsou v korpusu SYN zastoupeny dostatečným počtem textů (tedy alespoň čtyřmi vědeckými monografiemi).

2.2.4 Přehled vlastností přiřazených trénovacím datům

Všem textovým pozicím v trénovacích datech jsou přidělovány hodnoty jednotlivých vlastností. Některé vlastnosti jsou přiřazeny ručně (zařazení mezi termíny či mezi víceslovné

[8]Pokud by pozice byly vybrány náhodně, pro výzkum terminologie by to představovalo zásadní problém. V neseřazeném textu by nebylo možno vyhledávat termínové kolokace (víceslovné termíny), protože by nebyly k dispozici asociační míry (víceslovné termíny by nebylo možné označovat dokonce ani ručně). Kromě toho je při vyhledávání termínů zapotřebí mít informaci o kontextu (viz kap. 2.2.4), které můžou při vyhledávání termínů hrát velkou roli.

[9]Důvodem pro výběr dat z různých textů v rámci oboru je snaha vyhnout se natrénování na jeden konkrétní text.

Tabulka 2.4: Seznam vlastností, které jsou přiřazeny jednotlivým textovým pozicím v trénovacích datech. Vlastnosti označené zkratkou MWT se používají pouze pro automatické vyhledávání víceslovných termínů. (SCI - subkorpus akademických textů, COMPAR - srovnávací subkorpus obecných textů)

Atribut	**Popis atributu**
T	ručně označený termín
MWT	ručně označená součást víceslovného termínu
RFQ_{disc}	relativní frekvence v disciplíně
RFQ_{sci}	relativní frekvence ve SCI
RFQ_{compar}	relativní frekvence v COMPAR
$\text{RFQ}_{disc}\text{RFQ}_{compar}$	poměr relativních frekvencí v disciplíně a v COMPAR
$\text{RFQ}_{sci}\text{RFQ}_{compar}$	poměr relativních frekvencí ve SCI a v COMPAR
NoCompar	instance se vůbec nevyskytuje v COMPAR
RDist	relativní distribuce v disciplínách
SDRFQ	standardní odchylka relativních frekvencí v disciplínách
ARF	průměrná redukovaná četnost
RARF	relativní průměrná redukovaná četnost
SDRD	směrodatná odchylka relativní vzdálenosti sousedních výskytů
Len_{syl}	délka slova ve slabikách
Struct	ne/obvyklost struktury slova
CaseU	lemma začíná velkým písmenem (vlastní jména)
KontextHH	kontext: Herfindahl-Hirschmanův index
KontextE	kontext: entropie
KontextHgen	vážený průměr relativních frekvencí předcházejícího kontextu
MWT:T1	terminologická platnost první pozice (pro víceslovné termíny)
MWT:T2	terminologická platnost druhá pozice (pro víceslovné termíny)
MWT:MI-score	asociační míra MI-score (pro víceslovné termíny)
MWT:t-score	asociační míra t-score (pro víceslovné termíny)
MWT:Oblig	asociační míra obligatornost (pro víceslovné termíny)
MWT:Prox	asociační míra proximita (pro víceslovné termíny)
MWT:Modus	modus výskytu (pro víceslovné termíny)
MWT:AFC5	absolutní frekvence kolokátu v kontextu -5 až 5 (pro víceslovné termíny)
MWT:AFC3	absolutní frekvence kolokátu v kontextu -3 až 3 (pro víceslovné termíny)

termíny), ostatní jsou zpracovávány automaticky pomocí počítačového programu[10]. Jednotlivé vlastnosti byly vybrány na základě dosavadních výzkumů (viz kap. 1.2.2.1) a zkušeností s automatickým vyhledáváním termínů (Šrajerová, 2009a,b; Šrajerová et al., 2009). Výsledkem je rozsáhlá tabulka o osmi tisících řádků (počet textových pozic v trénovacích datech) a několika desítkách sloupců s hodnotami atributů[11].

Tabulka 2.4 poskytuje přehled všech vlastností použitých při výzkumu. Popisy vlastností (viz níže) se snaží vystihnout jednotlivé rysy co nejstručněji[12]. Pokud je to nutné, přiřazuji v příloze E jednotlivým rysům vzorce, podle kterých se vypočítá jejich hodnota, aby bylo možné zopakovat všechny experimenty. Vlastnosti označené jako MWT (multi-word term) se používají při vyhledávání víceslovných termínů, ostatní se týkají pouze termínů jednoslovných.

Není úkolem následujícího výčtu hodnotit užitečnost nebo důležitost jednotlivých vlastností, ale nashromáždit a popsat velké množství rysů, z nichž si data-miningové nástroje budou vybírat. Ty jsou totiž schopny vybrat z předložených rysů ty, které mají největší vliv na proces učení a na úspěšnost výsledků daného výzkumu (vyhledávání termínů). Toto hodnocení vlastností se nazývá feature ranking nebo feature selection a jsou mu věnovány kapitoly 5.1.1 a 5.2.1. Právě feature ranking ukáže, kterými rysy se více zabývat a popisovat jejich roli v automatickém vyhledávání termínů.

T (Termín) Tento atribut může mít dvě hodnoty, 1 nebo 0, přičemž 1 znamená „termín“ a 0 znamená „netermín“. Jako termíny jsou ručně označeny všechny textové pozice (obv. slova), které se buď vyskytují v terminologickém slovníku dané disciplíny jako samostatná hesla (tedy nikoli pouze jako součást víceslovných termínů), nebo které byly označeny za termíny odborníkem v daném oboru (viz kap. 2.2.5), případně i monosémní slova příbuzná s takto potvrzeným termínem (viz tamtéž).

Jako netermín jsou označeny textové pozice, které jsou obvyklými slovy běžně užívanými v neakademických textech (běžné spojky a předložky, pomocná slovesa, zájmena a další). V akademických textech se často vyskytují tzv. oborově nespecifické vědecké termíny (viz 1.3.1.2) - ty jsou také označeny jako netermíny.

[10] Jde o program vytvořený speciálně pro tento výzkum, jeho autorem je O. Kovářík.

[11] Ve skutečnosti jsou vytvořeny dvě sady dat, tedy dvě stejně rozsáhlé matice, jedna pro slovní tvary a jedna pro lemmata. Součástí experimentů totiž je i zjišťování, jak odlišně pracují data-miningové nástroje s lemmaty a se slovními tvary (více k tomu viz 2.2.6).

[12] Většina předkládaných vlastností vychází buď částečně, nebo zcela z rysů použitých v předchozích výzkumech terminologie, kontextu atp. (viz kap. 1.2.2).

N (Nejasné) Jako nejasné jsou ručně označeny textové pozice, které nelze jednoznačně zařadit mezi termíny nebo netermíny[13] (viz 2.2.5). Dále do této kategorie patří pozice, které by mohly být příčinou nesprávného naučení metody automatického vyhledávání. Jde o číslice, slova z jiného jazyka, slova obsahující překlepy apod. Všechny textové pozice označené jako nejasné jsou z trénovacích dat vyřazeny po dobu učícího procesu.

Tento atribut může mít dvě hodnoty, 1 nebo 0, přičemž 1 znamená „nejasné“ a 0 znamená „není nejasné“.

MWT (Víceslovný termín) Zkratkou MWT (Multi-Word Term) jsou ručně označeny všechny textové pozice (obv. slova), které jsou součástí víceslovného termínu a jsou v textu umístěny v bezprostředním sousedství[14]. Za víceslovný termín jsou označeny všechny součásti víceslovných termínů, ať už se jedná o termíny, nebo ne (víceslovné termíny obvykle obsahují alespoň jeden jednoslovný termín, ale existují i víceslovné termíny složené pouze z netermínů nebo pouze z termínů[15]).

Tento atribut může mít dvě hodnoty, 1 nebo 0, přičemž 1 znamená „součást víceslovného termínu“ a 0 znamená „není součástí víceslovného termínu“.

RFQ_{disc} (Relativní frekvence v disciplíně) Hodnota relativní frekvence v disciplíně se vypočítá jako poměr frekvence výskytů textové pozice v textech daného oboru ku počtu všech textových pozic v oboru.

Vzorec pro výpočet tohoto atributu je uveden v příloze E.

RFQ_{sci} (Relativní frekvence v subkorpusu SCI) Poměr frekvence výskytů textové pozice v textech subkorpusu SCI (akademické texty) ku počtu všech textových pozic v subkorpusu SCI.

Vzorec pro výpočet tohoto atributu je uveden v příloze E.

[13] Některým instancím byly přiřazeny hodnoty atributů T i N 1 - to jsou instance, u kterých se předpokládá, že by mohlo jít o termín, ale tento předpoklad nebylo možné potvrdit za pomoci terminologického slovníku ani za pomoci konzultace s odborníkem z oboru.

[14] Předkládaný výzkum není zaměřen na vyhledávání víceslovných termínů, jejichž jednotlivé prvky jsou od sebe v textu odděleny např. závorkami nebo spojkami.

[15] Existují i jiné typy termínů složené například ze slov i číslic, od těch ale v předkládaném výzkumu odhlížíme.

RFQ$_{compar}$ (Relativní frekvence v subkorpusu COMPAR) Poměr frekvence výskytů textové pozice v textech subkorpusu COMPAR ku počtu všech textových pozic v subkorpusu. Může se stát, že se daná textová pozice v subkorpusu COMPAR nevyskytuje vůbec, potom je jí přidělena hodnota 0, a atributu NoCompar (viz níže) hodnota 1.

Vzorec pro výpočet tohoto atributu je uveden v příloze E.

RFQ$_{disc}$RFQ$_{compar}$ (Poměr relativních frekvencí disciplína/obecné texty) Poměr relativní frekvence instance v dané disciplíně ku relativní frekvenci téže instance v subkorpusu COMPAR. Zjednodušeně řečeno jde o to, jak často se textová pozice vyskytuje v daném oboru oproti subkorpusu neakademických textů. Pokud se vůbec nevyskytuje v subkorpusu COMPAR, atributu RFQ$_{disc}$RFQ$_{compar}$ je přiřazena speciální hodnota, která se odvíjí od maximální nabyté hodnoty[16]. Pro tyto případy je vytvořen doplňkový atribut NoCompar (viz níže).

Čím vyšší je hodnota atributu, tím větší je nepoměr mezi frekvencí v textech daného oboru a frekvencí v textech neakademických (ve prospěch textů oboru).

Vzorec pro výpočet tohoto atributu je uveden v příloze E.

RFQ$_{sci}$RFQ$_{compar}$ (Poměr relativních frekvencí odborné texty/obecné texty) Poměr relativní frekvence instance v akademických a neakademických textech.

Pokud se daná textová pozice vůbec nevyskytuje v subkorpusu COMPAR, atributu RFQ$_{sci}$RFQ$_{compar}$ je přiřazena speciální hodnota, která se odvíjí od maximální nabyté hodnoty[17]. Pro tyto případy je vytvořen doplňkový atribut NoCompar (viz níže).

Čím vyšší je hodnota atributu, tím větší je nepoměr mezi frekvencí v akademických a neakademických textech (ve prospěch akademických textů).

Vzorec pro výpočet tohoto atributu je uveden v příloze E.

NoCompar (Nulový výskyt v subkorpusu COMPAR) Doplňkový atribut k poměru relativních frekvencí. Pokud se daná textová pozice vyskytuje v subkorpusu akademických textů, ale má nulovou frekvenci v subkorpusu COMPAR, je hodnota atri-

[16]Ve skutečnosti v tomto případě nabývá atribut hodnoty nekonečno. To zkresluje celkové výsledky, proto se z technických důvodů těmto pozicím přiřazuje zvláštní hodnota, která je jasně odliší od ostatních instancí, ale výsledky přitom nezkreslí.

[17]Ve skutečnosti v tomto případě nabývá atribut hodnoty nekonečno. To zkresluje celkové výsledky, proto se z technických důvodů těmto pozicím přiřazuje zvláštní hodnota, která je jasně odliší od ostatních instancí, ale výsledky přitom nezkreslí.

butu NoCompar 1, ve všech ostatních případech je hodnota 0. Tento atribut souvisí s atributem RFQ_{compar}.

RDist (Relativní distribuce) Základem pro výpočet hodnoty relativní distribuce je počet oborů v korpusu SYN2010 v rámci textového typu SCI, v kterých se daná textová pozice vyskytuje. Tento počet je vydělen počtem všech akademických oborů v subkorpusu SCI, atribut tedy nabývá hodnoty v intervalu 0-1, což umožňuje reprodukovatelnost experimentů v případě jakéhokoli množství oborů.

Vzorec pro výpočet tohoto atributu je uveden v příloze E.

SDRFQ (Směrodatná odchylka RFQ_{disc}) Směrodatná odchylka relativní frekvence dané textové pozice ve všech oborech. Jde o údaj vypovídající jiným způsobem o distribuci pozice v různých oborech. Oproti relativní distribuci ale zohledňuje i frekvenci výskytů textové pozice, a tedy i to, ve kterém oboru je dané slovo pevněji zakotveno[18].

Čím menší je hodnota směrodatné odchylky, tím je frekvence textové pozice v oborech rovnoměrnější.

Vzorec pro výpočet tohoto atributu je uveden v příloze E.

ARF (Průměrná redukovaná četnost) Průměrná redukovaná četnost vypovídá o pravidelnosti či rovnoměrnosti rozmístění slova v korpusu a zároveň o jeho frekvenci. Čím je textová pozice frekventovanější a zároveň rovnoměrněji rozložena, tím je hodnota ARF vyšší.

Korpus je pro výpočet hodnoty ARF rozdělen do úseků, jejichž počet odpovídá celkové frekvenci dané pozice (pokud se tedy slovo v korpusu vyskytuje třikrát, korpus se rozdělí na tři stejně velké úseky, pokud je frekvence slova 1000, rozdělí se korpus na 1000 úseků). Hodnota ARF se rovná počtu úseků, ve kterých se slovo alespoň jednou nachází[19].

Hodnoty ARF pro jednotlivá slova jsou převzaty z korpusu SYN2010[20].

[18] Může se například stát, že relativní distribuce slova *obsah* se bude blížit hodnotě 1, protože slovo se vyskytne ve většině akademických oborů. V jednom z oborů (ve fyzice) ale jeho relativní frekvence prudce stoupne oproti ostatním oborům, a tak se ukáže, že je pevněji zakotveno právě v této disciplíně (je zde termínem, kdežto v ostatních oborech ne).

[19] Je přitom třeba vypočítat redukovanou četnost od všech možných začátků v rámci prvního úseku, čímž se vyhneme tomu, že by výsledky zkreslovalo pevné umístění hranic mezi úseky.

[20] Ve skutečnosti se hodnota ARF pro korpus SYN2010 počítá způsobem, který aproximuje zde popsaný postup (Savický, Hlaváčová, 2003).

RARF (Relativní průměrná redukovaná četnost) Relativní průměrná redukovaná četnost je ARF vydělená frekvencí slova v korpusu. Oproti průměrné redukované četnosti potlačuje silný vliv frekvence. Prostřednictvím ARF můžeme srovnávat pouze slova s podobnou frekvencí, prostřednictvím RARF lze srovnávat i textové pozice frekvenčně velmi odlišné.

Tento atribut nabývá hodnoty v intervalu 0-1. Čím vyšší je hodnota RARF, tím rovnoměrněji jsou jednotlivé výskyty rozmístěné v korpusu.

Vzorec pro výpočet tohoto atributu je uveden v příloze E.

SDRD (Směrodatná odchylka relativní vzdálenosti sousedních výskytů) Podobně jako ARF, jemuž se SDRD do určité míry podobá, se hodnoty vypočítávají pro celý korpus SYN2010. Jde o relativní odchylku vzdáleností mezi jednotlivými výskyty vydělených frekvencí daného slova. Smyslem SDRD je zjistit, jak pravidelně jsou jednotlivé výskyty rozmístěny v korpusu.

Čím vyšší je hodnota tohoto atributu, tím menší je rovnoměrnost rozmístění slova v korpusu (jednotlivé výskyty jsou ve shlucích).

Vzorec pro výpočet tohoto atributu je uveden v příloze E.

Len$_{syl}$ (Délka slova ve slabikách) Hodnotou tohoto atributu je počet slabik dané textové pozice. Slabiky ve slovech se počítají podle speciálních pravidel, která zachycují možné struktury českých slabik[21].

Struct (Struktura slova) Atribut zachycuje ne/obvyklost formy slova. Mnohé termíny jsou cizího původu a jejich struktura je v rámci českého textu často poměrně nápadná (viz kap. 1.2.2.2). Hodnota tohoto atributu se vypočítá jako součet pravděpodobností každého bigramu ve slově. Pravděpodobnosti bigramů jsou převzaty z korpusu SYN2010.

Čím je hodnota atributu nižší, tím je forma slova neobvyklejší.

CaseU (Lemma začíná vždy velkým písmenem) Pokud lemma dané instance začíná pouze velkým písmenem, atribut má hodnotu 1, v ostatních případech má hodnotu 0.

[21] Autorem programu pro automatické počítání slabik je V. Cvrček.

KontextHH (Herfindahl-Hirschmanův index) Herfindahl-Hirschmanův index měří diverzitu jevů a zároveň i rovnoměrnost zastoupení jevů ve zkoumaném kontextu, hodnotí tedy počet typů v kontextu slova, stejně jako zastoupení jednotlivých typů. Mezi atributy KontextHH a KontextE je silná korelace (hodnota korelace je -0,7; viz tab. 5.2 v kap. 5.1.2), bylo by tedy možné použít jen jeden z nich.

Hodnota tohoto atributu se pohybuje v intervalu 0-1, přičemž čím vyšší je hodnota, tím je kontext uspořádanější.

Vzorec pro výpočet tohoto atributu je uveden v příloze E.

KontextHHp1 (Herfindahl-Hirschmanův index, pravý kontext) Atribut hodnotí počet a zastoupení typů v bezprostředním pravém kontextu slova. V rámci textu se může měřit levý i pravý kontext, přičemž pravý kontext je obecně méně variabilní (poskytuje tedy o dané textové pozici více informací) (Cvrček, 2013).

Hodnota tohoto atributu se pohybuje v intervalu 0-1, přičemž čím vyšší je hodnota, tím je pravý kontext uspořádanější.

KontextHHl1 (Herfindahl-Hirschmanův index, levý kontext) Atribut hodnotí počet a zastoupení typů v bezprostředním levém kontextu slova.

Hodnota tohoto atributu se pohybuje v intervalu 0-1, přičemž čím vyšší je hodnota, tím je levý kontext uspořádanější.

KontextE (Entropie) Entropie je míra neuspořádanosti, v našem případě míra neuspořádanosti kontextu. Mezi atributy KontextHH a KontextE je silná korelace (hodnota korelace je -0,7, viz tab. 5.2), bylo by tedy možné použít jen jeden z nich.

Čím nižší je hodnota tohoto atributu, tím je kontext uspořádanější.

Vzorec pro výpočet tohoto atributu je uveden v příloze E.

KontextEp1 (Entropie pravého kontextu) Míra neuspořádanosti bezprostředního pravého kontextu. Pravý kontext je obecně méně variabilní, poskytuje tedy o dané textové pozici více informací (Cvrček, 2013).

Čím nižší je hodnota tohoto atributu, tím je kontext uspořádanější.

KontextEl1 (Entropie levého kontextu) Míra neuspořádanosti bezprostředního levého kontextu.

Čím nižší je hodnota tohoto atributu, tím je kontext uspořádanější.

Hgen (Vážený průměr relativních frekvencí předcházejícího kontextu) Atribut Hgen představuje vážený průměr relativních frekvencí slov v kontextu bezprostředně předcházejících zkoumanému slovu. Počítá se se dvěma až pěti předcházejícími slovy a podle toho se atribut značí Hgen2 až Hgen5.

Čím méně frekventovaná slova se v kontextu předcházejícím zkoumanému slovu vyskytují, tím je hodnota atributu vyšší.

Vzorec pro výpočet tohoto atributu je uveden v příloze E.

MWT:T1 (Terminologická platnost prvního slova) Jde o atribut používaný pouze při vyhledávání víceslovných termínů. Velká část víceslovných termínů obsahuje alespoň jeden jednoslovný termín, v některých případech i několik jednoslovných termínů. Proto je možné využít automatického přidělení hodnoty terminologické platnosti jednoslovným termínům v prvním kole vyhledávání.

Atribut nabývá hodnot mezi 0 a 1 podle toho, zda je první pozice ve zkoumaném bigramu termínem (o tom se rozhoduje na základě automatické identifikace jednoslovných termínů).

MWT:T2 (Terminologická platnost druhého slova) (viz MWT:T1)

Atribut nabývá hodnot mezi 0 a 1 podle toho, zda je druhá pozice ve zkoumaném bigramu termínem (o tom se rozhoduje na základě automatické identifikace jednoslovných termínů).

MWT:MI-score (Asociační míra MI-score) Asociační míra MI-score (Mutual Information, vzájemná informace) je atribut používaný pouze při vyhledávání víceslovných termínů. Měří vzájemnou závislost dvou náhodných proměnných; v korpusové lingvistice je MI-score využíváno k vyhledávání kolokací - jde o stanovení pravděpodobnosti, s jakou se jedno slovo vyskytne v kontextu druhého.

MI-score se měří pro všechny bigramy v trénovacích datech.

Vzorec pro výpočet tohoto atributu je uveden v příloze E.

MWT:t-score (Asociační míra t-score) Asociační míra t-score je atribut používaný pouze při vyhledávání víceslovných termínů. V korpusové lingvistice je t-score využíváno k vyhledávání kolokací. Vychází ze statistické metody testování hypotéz pomocí tzv. t-testu. Testuje se, jak hodně výskyty dvojic slov (po sobě jdoucích textových pozic) odpovídají náhodnému rozložení slov v korpusu.

T-score se měří pro všechny bigramy v trénovacích datech.

Vzorec pro výpočet tohoto atributu je uveden v příloze E.

MWT:Prox (Proximita) Proximita je atribut používaný pouze při vyhledávání víceslovných termínů . Jde o průměr absolutních hodnot textových vzdáleností dvou slov (počítáno v rozsahu tří pozic v levém kontextu až tří pozic v pravém) (Cvrček, 2013).

Atribut nabývá hodnot 1 až 3 pro slova, která se setkávají v blízkém kontextu, přičemž hodnota 1 se vyskytuje u kombinace slov sousedících výhradně na bezprostředních pozicích. Čím vyšší hodnota, tím vzdálenější jsou pozice, na kterých se slova vyskytují. Další možná hodnota je 0, ta je vyhrazena pro slova, která se spolu v blízkém kontextu (pozice -3 až 3) nevyskytují.

MWT:Oblig (Vzájemná obligatornost kontextu) Vzájemná obligatornost je atribut používaný pouze při vyhledávání víceslovných termínů. Je to podmíněnost vzájemného souvýskytu dvou členů bigramu. Sleduje se počet výskytů členu B bigramu v blízkém kontextu členu A a naopak. Tento počet je vyjádřen jako procento z celkové frekvence členu A (resp. B). Vyšší hodnota je vybrána jako vzájemná obligatornost kontextu obou členů bigramu (Cvrček, 2013).

Atribut nabývá hodnot 0 až 100 % - nulová hodnota znamená, že se slova v bigramu nevyskytují ve stejném kontextu, 100 % ukazuje na situaci, kdy je alespoň jedno slovo vždy realizováno spolu s druhým slovem bigramu.

MWT:Modus (Modus výskytu) Modus výskytu je atribut používaný pouze při vyhledávání víceslovných termínů. Jde o kontextovou pozici, na níž se jeden člen bigramu nachází nejčastěji vzhledem k druhému členu (Cvrček, 2013).

Hodnota vlastnosti je vyjádřena číslem podle toho, jak jsou od sebe typicky členy bigramu v textu vzdáleny.

MWT:AFC5 (Absolutní frekvence kolokátu v kontextu -5 až 5) AFC5 je atribut používaný pouze při vyhledávání víceslovných termínů. Je to počet souvýskytů kolokátu a KWICu v kontextu o rozsahu pět pozic vlevo a pět pozic vpravo (Cvrček, 2013).

MWT:AFC3 (Absolutní frekvence kolokátu v kontextu -3 až 3) AFC3 je atribut používaný pouze při vyhledávání víceslovných termínů. Je to počet souvýskytů kolokátu a KWICu v blízkém kontextu (tedy v kontextu o rozsahu tři pozice vlevo a tři pozice vpravo) (Cvrček, 2013).

2.2.5 Ruční vyhledání termínů

Ručním označováním termínů začíná proces přidělování hodnot jednotlivým atributům v rámci trénovacích dat. Není to úkol ani zdaleka jednoduchý, protože v konkrétních textech se termíny chovají do značné míry jinak než v teoretických popisech (viz kap. 1.2.1). Přesnost ručního označování ovlivňuje učící proces, a tedy i výslednou metodu automatického vyhledávání. Zároveň má vliv i na hodnocení důležitosti vlastností termínů, proto je třeba věnovat mu velkou pozornost.

V trénovacích datech byly označovány zvlášť jednoslovné termíny a zvlášť termíny víceslovné. Jednoslovnými termíny se zde míní samostatně stojící, i takové, které jsou součástí víceslovných termínů.

Víceslovné termíny obvykle obsahují jeden i více jednoslovných termínů, v některých případech mohou být složeny pouze z netermínů. Pro účely našeho výzkumu byly za víceslovné označovány jen takové termíny, jejichž jednotlivé prvky jsou v textu v těsné blízkosti (přímo sousedí). Většina ručně nalezených víceslovných termínů je dvouslovná, nejdelší víceslovné termíny v trénovacích datech se skládají z pěti textových pozic.

2.2.5.1 Metodika ručního označování textových pozic

Původním předpokladem pro ruční označování bylo, že v textech budou označeny pouze ty termíny, které lze najít v terminologických slovnících příslušných oborů. Ukázalo se ale, že práce s terminologickými slovníky je problematická, a to především ze dvou důvodů:

1) Slovníky nejsou určeny pro účely lingvistického výzkumu oborové terminologie, ale pro odborníky daného oboru, studenty, případně zájemce o daný obor. Jsou tedy

zaměřeny spíše na vysvětlení termínů, které jsou těmto uživatelům neznámé, než na vytvoření souhrnu všech termínů v oboru (to by navíc bylo v mnoha oborech takřka nemožné). Ve slovnících tedy leckdy chybí termíny, které jsou užívané v konkrétních textech, například mnohé víceslovné termíny (ve slovníku jsou uvedeny jen jednotlivé složky možných kombinací). Naopak často najdeme termíny z jiného oboru, jejichž znalost je pro výzkum v daném oboru užitečná: např. ve Velkém sociologickém slovníku (1996) jsou i termíny z oboru statistiky, která je součástí sociologické metodologie.

2) Terminologie konkrétních oborů jsou z velké většiny tvořeny substantivy. To s sebou přináší i tendenci považovat za termín substantiva, ale méně už příbuzná adjektiva nebo slovesa. Přitom v mnoha případech (i když určitě ne ve všech) mají slova příbuzná se substantivním termínem vysokou terminologickou platnost. Mnohdy se tedy stane, že ve slovníku je zahrnuto pouze substantivum, nikoli už slova příbuzná. Příkladem může být termín *revizionismus* (SOC), který v terminologickém slovníku najdeme, a termín *revizionistický* (SOC), který už ve slovníku zahrnut není (více v kap. 2.2.5.2).

Z toho důvodu bylo při přípravě dat nutné konzultovat s odborníky[22] daných oborů, kteří jsou obvykle schopni rozhodnout, zda je určité slovo nebo kolokace termínem. Díky těmto konzultacím byla většina textových pozic v trénovacích datech s jistotou zařazena mezi netermíny, nebo mezi termíny, případně víceslovné termíny. Přesto se vyskytla slova nebo kolokace, jejichž jednoznačné zařazení mezi termíny nebo netermíny nebylo možné. Proto byly pro ruční označování vytvořeny celkově tři kategorie (viz tab. 2.5):

1. jednoslovné termíny (s hodnotami termín a netermín)
2. nejasné, tj. takové, o nichž není možné s jistotou rozhodnout, zda patří mezi termíny, či netermíny (s hodnotami nejasné a není nejasné)[23]
3. víceslovné termíny (s hodnotami je či není součástí víceslovného termínu).

Textové pozice přidělené do kategorie nejasné jsou při učícím procesu z trénovacích dat vyřazeny, protože by mohly negativně ovlivnit natrénování metody.

[22]Poděkování patří ing. O. Kováříkovi (obor informatika), Mgr. O. Hanusovi (obor literární věda), Mgr. P. Pabinovi, PhD (obor sociologie) a MUDr. R. Čáberovi (obor medicína).

[23]Mj. zde jde i o záležitost neustálého vývoje, během nějž se některé termíny stávají pevnější součástí systému termínů určité disciplíny.

Tabulka 2.5: Příklady výsledků ručního označování ve čtyřech zkoumaných disciplínách. Pokud je textová pozice termínem, je v odpovídajícím sloupci označena jako 1. Pokud jde o netermín, je v tomtéž sloupci označena jako 0. Pokud nelze jednoznačně rozhodnout, zda jde o termín, či netermín, je textové pozici ve sloupci „Nejasné“ přiřazena hodnota 1. Pokud je textové pozici přiřazena hodnota 1 v prvním i druhém sloupci, jedná se pravděpodobně o termín, ale nelze to potvrdit např. na základě terminologického slovníku. Textové pozice, které jsou součástí víceslovných termínů, jsou v posledním sloupci označeny jako 1.

Textová pozice	**Termín**	**Nejasné**	**Víceslovný t.**	**Textová pozice**	**Termín**	**Nejasné**	**Víceslovný t.**
COM				**LIT**			
terminálového	1	0	1	autor	1	0	0
programu	1	0	1	,	0	0	0
.	0	0	0	jenž	0	0	0
Jelikož	0	0	0	konstruuje	1	1	0
přímé	0	1	1	text	1	0	0
adresování	1	0	1	a	0	0	0
předpokládá	0	0	0	fikční	1	0	1
přímé	0	1	0	svět	0	0	1
zadávání	1	1	0	textu	1	0	0
adresy	1	0	0	a	0	0	0
,	0	0	0	čtenář	1	0	0
není	0	0	0	,	0	0	0
bohužel	0	0	0	jenž	0	0	0
možné	0	0	0	čte	0	1	0
adresu	1	0	0	text	1	0	0
přijmout	1	1	0	a	0	0	0
a	0	0	0	rekonstruuje	1	1	0
zapsat	1	1	0	jeho	0	0	0
ji	0	0	0	fikční	1	0	1
do	0	0	0	svět	0	0	1
registru	1	0	0	.	0	0	0
MED				**SOC**			
kardiologie	1	0	0	Empiricky	1	1	0
,	0	0	0	chceme	0	0	0
kde	0	0	0	zkoumat	0	0	0
hlavními	0	0	0	společnost	1	0	0
doménami	0	0	0	proto	0	0	0
je	0	0	0	,	0	0	0
vyšetřování	1	0	0	protože	0	0	0
stavu	0	1	0	nám	0	0	0
ANS	1	0	0	nestačí	0	0	0
u	0	0	0	apriorní	0	0	0
nemocných	1	0	0	vědění	1	0	0
s	0	0	0	o	0	0	0
různými	0	0	0	společnosti	1	0	0
formami	0	0	0	ve	0	0	0
ischemické	1	0	1	formě	0	0	0
choroby	1	0	1	zdravého	0	1	1
srdeční	1	0	1	rozumu	0	1	1
,	0	0	0	.	0	0	0
zejména	0	0	0	Všichni	0	0	0
z	0	0	0	sociologové	1	0	0
hlediska	0	0	0	věří	0	0	0
jejich	0	0	0	,	0	0	0
prognózy	1	0	0	že	0	0	0

2.2.5.2 Nejčastější problémy při ručním vyhledávání termínů

Následující přehled nejčastějších problémů, se kterými se setká každý, kdo ručně označuje termíny v reálných textech, je spíše souhrnem poznámek svědčících o tom, jak můžou být reálná data vzdálená od teoretických úvah o jazyce. Řešením je stanovit si pravidla, podle kterých se označování termínů bude v daném výzkumu konzistentně řídit. Každý zde zmíněný problém je opatřen příklady a následuje popis řešení, pro které jsem se v rámci tohoto konkrétního výzkumu rozhodla.

Polysémie

Největším problémem při ručním označování termínů je polysémie. V ideálním případě by podle teoretických popisů měl termín být alespoň v rámci jednoho oboru monosémní a nezávislý na kontextu (viz kap.1.2.1.2). Při práci s konkrétními odbornými texty ale brzo zjistíme, že nejsou vzácné případy, kdy má slovo v rámci oboru (ale dokonce i v rámci jednoho textu) v různých kontextech různé významy, terminologické i neterminologické. Polyfunkčnost (mezioborová polysémie) je jevem ještě častějším (viz 1.3.1.5). Příkladem může být slovo *špičkový*[24], které je v energetice termínem (*špičkové zatížení, špičkový výkon* = zatížení/výkon ve špičce), ale je užíváno i neterminologicky (*špičkový odborník, špičkový manažer, špičková úroveň*), a to i v rámci krátkého úseku textu.

Řešení tohoto problému je následující: pokud je v rámci trénovacích dat jednoho oboru dané slovo alespoň jednou termínem, budou i ostatní výskyty slova považovány za termín (i kdyby šlo o neterminologické užití).

Pokud je dané slovo v jednom oboru termínem a v ostatních netermínem, zůstává tento rozdíl zachován.

Oborově nespecifické vědecké termíny

Důležitou roli v akademických textech hrají oborově nespecifické termíny (viz 1.3.1.2). Ty jsou sice často primárně termínem v nějakém oboru (obvykle ve filozofii, ale také třeba ve fyzice nebo ve statistice), ale s poměrně vysokou frekvencí se vyskytují v odborných textech velkého množství dalších oborů se sníženou terminologickou platností. Příkladem můžou být slova *funkce, systém, struktura, pojem, kategorie, data, analyzovat, interpretovat.*

[24]Podobné případy se vyskytují i v disciplínách, z nichž jsou vybrána trénovací data, např. *webové* ***stránce*** vs. *jak po* ***stránce*** *hardwaru, tak i programů* (COM) nebo *ve vedlejší komické* ***roli*** vs. *podstatnou* ***roli*** *přitom hrají nezáměrné významy* (LIT).

Podobné případy jsou v rámci tohoto výzkumu při ručním vyhledávání obvykle hodnoceny jako netermíny (pokud ovšem nepatří do terminologie daného oboru). Takto jednoduše se ale nelze rozhodnout vždy. Často rozhodnutí komplikuje fakt, že slovo je součástí víceslovného termínu a je tedy pociťováno jako termín (viz níže): *aproximovat libovolnou spojitou* ***funkci*** (COM), *estetická* ***funkce*** (LIT), *urogenitální* ***systém*** (MED). Pokud skutečně není možné jednoznačně rozhodnout, zda jde o termín či netermín, je třeba takovou textovou pozici zařadit do kategorie „nejasné“ (viz 2.2.5.1).

Terminologická platnost částí víceslovných termínů

Ve víceslovném termínu může vedle sebe stát několik termínů, ale častěji jsou kombinací termínů a netermínů (vč. oborově nespecifických termínů). Výjimečně je víceslovný termín tvořen pouze netermíny.

Problémem pro ruční označování termínů je tendence vnímat vyšší terminologickou platnost jednotlivých složek víceslovného termínu, než by byla pociťována u samostatně stojícího slova, zvlášť pokud se jedná o oborově nespecifický termín (viz i výše). Např. v literární vědě je velmi těžké rozhodnout, zda slova *funkce, norma, hodnota* ve víceslovných termínech *estetická funkce, estetická norma, estetická hodnota* jsou samostatné termíny. Pokud skutečně není možné jednoznačně rozhodnout, zda jde o termín či netermín, je třeba takovou textovou pozici zařadit do kategorie „nejasné“ (viz 2.2.5.1).

Odvozování

V terminologických slovnících jsou nejčastěji uváděna pouze podstatná jména, případně i přídavná jména, která se obvykle vyskytují jen jako součásti víceslovných termínů a nemají tedy samostatné heslo. Slovesa, natož pak další slovní druhy, lze ve slovnících nalézt jen zřídka. Přesto ale nelze říct, že by jiné slovní druhy než podstatná jména nemohly být termíny. Slovesa *skloňovat* a *časovat* nebo adjektiva *skloňovaný* a *časovaný* jsou v lingvistice stejně tak termíny, jako substantiva *skloňování* a *časování*.

V reálných odborných textech se vyskytuje mnoho slov příbuzných termínům, u nichž nemůžeme v rámci stanovené metodologie (za termín se označují taková slova, která jsou obsažena v terminologickém slovníku) jednoznačně určit, zda jde o termíny. Řešením je posoudit kromě příbuznosti s termínem také polysémnost zkoumané textové pozice - čím méně je polysémní, tím snáze lze zařadit mezi termíny. Například adjektivum *chlopenní*, odvozené od substantivního termínu *chlopeň* (MED), je zařazeno mezi termíny, na rozdíl

Tabulka 2.6: Množství ručně vyhledaných jednoslovných a víceslovných termínů v trénovacích datech (v procentech). Trénovací data obsahují celkově dva tisíce textových pozic pro každou ze zkoumaných disciplín. Za jednoslovné termíny jsou zde považovány pouze termíny stojící samostatně (nejsou součástí víceslovného termínu), u víceslovných termínů se započítávají všechny textové pozice, které obsazují.

Typ termínů (v %)	COM	LIT	MED	SOC
jednoslovné termíny	12.3	6.8	10.8	5.9
víceslovné termíny (textové pozice)	11.8	4.5	15.8	10.2
celkem termínů (textové pozice)	24.1	11.2	26.6	16.1

od adjektiva *zdravý*[25] příbuzného se substantivním termínem *zdraví*(SOC).

Formálně neobvyklé textové pozice

Většina akademických textů v korpusu SYN2010 obsahuje textové pozice, které jsou nějakým způsobem neobvyklé, jako např. číslice, slova z cizího jazyka (název knihy apod.), slova spojená pomlčkou či obsahující překlepy. Všechny tyto případy jsou z trénovacích dat vyřazeny (jsou označeny jako „nejasné“, viz 2.2.5.1), protože by mohly ovlivnit učící proces (jejich vlastnosti by mohly být mylně přisouzeny termínům).

2.2.5.3 Srovnání zkoumaných oborů

Na základě ručně označených dat lze porovnávat zkoumané obory (informatiku, literární vědu, lékařství a sociologii) co do počtu vyhledaných jednoslovných a víceslovných termínů (viz tab. 2.6). Každý obor má navíc svoje specifika, která při ručním vyhledávání hrála velkou roli a která se určitým způsobem podílejí i na úspěšnosti data-miningových metod při vyhledávání automatickém.

Kromě ručně označených termínů a netermínů je ve všech textech cca 18 % textových pozic obsazených interpunkcí. Zbytek textových pozic, nejvýš však 10 %, tvoří slova zařazená do kategorie „nejasné“ (vč. číslic, slov z cizího jazyka, překlepů apod.).

Každý z akademických oborů je specifický, a to jak počtem jednoslovných a víceslovných termínů, tak i tím, jak do nich pronikají termíny ze souvisejících oborů a do/z neakademických textů, například publicistiky.

[25] V trénovacích datech (SOC) se toto adjektivum vyskytuje v kolokaci *zdravý rozum*.

Informatika (COM)

Informatika je technický obor, mnoho termínů sdílí s ostatními technickými obory (*technologie, přístroj, konstruovat, zátěž, chod*), ale také s obory, jako je matematika nebo fyzika (*rychlost, zrychlení, platnost, pravděpodobnost, algoritmus*). Stejně jako v ostatních technických oborech, i v informatice je poměrně velké množství termínů (cca čtvrtina trénovacích dat je tvořena termíny, zhruba stejně jednoslovnými jako víceslovnými, viz tab. 2.6). Specifikem tohoto oboru je rozšířenost jeho terminologie do neakademických textů způsobená všeobecným každodenním užíváním výpočetní techniky. Znalost některých vysoce odborných termínů mezi širokou veřejností je dnes už samozřejmá, i když porozumění termínům bývá často jen povrchní (*IP adresa, webový prohlížeč, webová stránka, program, procesor*). Právě tím, že se termíny s původně vysokou terminologickou platností používají se změněným (rozšířeným, zúženým, posunutým) významem, se jejich terminologičnost v neakademických textech snižuje (viz kap. 1.3.1.1). Ruční vyhledávání termínů v tomto oboru znesnadňuje terminologizace některých běžně užívaných slov (*zadávání* ***adresy***, ***příjem*** *zpráv,* ***připojení*** *k internetu*) nebo výše zmíněné sdílení termínů s dalšími obory, zvláště technickými (např. otázka, zda je *přístroj* termínem v informatice).

Literární věda (LIT)

Počet termínů v textech oboru literární věda je mnohem nižší než v případě informatiky (viz tab. 2.6), přičemž většinu textových pozic obsazených termíny tvoří samostatně stojící termíny jednoslovné. Ve zkoumaných literárněvědných textech je nejmenší poměr víceslovných termínů ku termínům jednoslovným. Literární věda je terminologicky propojena s ostatními uměnovědami, filozofií a částečně i s lingvistikou (*umění, umělecké dílo, estetika, symbolika, anticko-evropské hodnotové konvence, text*). Největším problémem při ručním vyhledávání termínů jsou slova, která by mohla být termínem, ale v textech jsou užívána neterminologicky, v obecnějším významu (*popis, přívlastek, smysl, forma, znak, výklad, synonymum*), a dále slova, u nichž není jednoduché rozhodnout, zda se jedná o termíny, či o oborově nespecifické termíny (*estetická norma, hodnota, funkce, tématický plán*).

Lékařství (MED)

Lékařství jako aplikovaná přírodovědná disciplína sdílí velké množství termínů s ostatními přírodovědnými obory, jako je biologie, chemie nebo fyzika (*reaktivita, genetické poruchy, kyslík, frekvence, tlak*). V textech tohoto oboru je vůbec nejvyšší množství termínů,

zvláště víceslovných (viz tab. 2.6). Ruční vyhledávání termínů v lékařských textech je nejjednodušší ze všech čtyř zkoumaných oborů - velkou roli v tom hraje i častá formální neobvyklost termínů cizího původu (*gastrointestinální, urogenitální, regurgitace pulmonální chlopně, transvalvulární gradient*).

Sociologie (SOC)

Sociologie je obor společenskovědní, ale nesdílí termíny pouze s dalšími společenskými vědami (politologie, ekonomie, psychologie, právo). Je totiž metodologicky, a tedy i terminologicky úzce propojen také se statistikou (*tvrdá data, objektivní měření, kvalitativní postupy, empirické zkoumání*). Statistické termíny jsou dokonce ve značné míře zahrnuty do sociologických terminologických slovníků. Ač tedy nejde o termíny čistě sociologické, jsou v předkládaném výzkumu v sociologických textech ručně označovány jako termíny (jeden z použitých textů z oboru sociologie se věnuje statistickému zpracování sociologických dat, obsahuje tedy velké množství statistických termínů). Ve zkoumaných sociologických textech je oproti lékařství a informatice poměrně malé množství termínů (viz tab. 2.6).

2.2.6 Výběr nejvhodnější formy trénovacích dat

V korpusovém výzkumu automatického vyhledávání termínů v českých textech[26] proti sobě stojí dva požadavky:

1) metoda vyhledávání termínů by měla mít **vysokou úspěšnost**, aby byly závěry o termínech a terminologii na ní založené co nejspolehlivější, a

2) měla by být v co nejvyšší míře **řízena korpusovými daty (corpus-driven)**[27]. Důvodů, proč je corpus-driven přístup žádoucí právě v rámci předkládané práce, je několik: a. tento přístup je spolehlivější ve svých závěrech, protože se opírá o skutečná data[28] (ač analýza dat i interpretace výsledků může být velmi náročná); b. výzkum řízený korpusovými daty může přinést zcela nové poznatky o jazykových jevech (termínech) (viz kap.1.1.3); c. z praktického hlediska je corpus-driven přístup v tomto případě výhodný proto, že nepotřebuje využívat lingvistického značkování korpusu vč. lemmatizace, a jeho výsledky jsou proto využitelné pro jakýkoli (tedy i neoznačkovaný) text.

[26]Kromě předkládané dizertační práce jde o předcházející výzkum. V jeho rámci bylo publikováno nebo prezentováno několik studií (Šrajerová, 2009a,b; Šrajerová et al., 2009).

[27]Pojmu corpus-driven přístup je věnována kap. 1.1.3.

[28]Skutečnými daty jsou v tomto případě reálné akademické texty, v nichž se termíny vyskytují, nikoli tedy seznamy termínů nebo popis toho, jak by termín měl vypadat.

Oba tyto požadavky je třeba vyvážit při rozhodování o konkrétní podobě materiálu, na jehož základě se budou provádět jednotlivé experimenty (např. hledání nejúspěšnější metody či vyhodnocování charakteristických rysů termínů). Pokud by corpus-driven přístup snížil úspěšnost vyhledávání termínů do té míry, že by nebylo možné se o výsledky opřít, pak by bylo nutné se ho vzdát. V opačném případě pak převáží jeho výhody (viz body a., b., c. výše).

Pro účely experimentů v kapitolách 4 a 5 je třeba se rozhodnout, jakou podobu budou textové pozice mít (lemmata či slovní tvary), jak budou seřazeny a které vlastnosti budou použity v experimentech. Na tomto základě lze formulovat tři následující otázky:

1. Je výhodnější použít data, v nichž se za sebou budou objevovat jednotlivé textové pozice tak, jak jdou za sebou v textu, nebo každé lemma/tvar brát v úvahu pouze jednou (**sloučit všechny výskyty** téhož do jediné instance, nebo je **opakovat**)?
2. Je vhodnější využít **lemmatizaci**, nebo pracovat s konkrétními **tvary slov**?
3. Je nutné využít **morfologického značkování** a přidělit jednotlivým textovým pozicím slovní druh?

Zodpovězením výše zmíněných otázek získáme oporu pro sestavení nejvhodnější formy materiálu pro další výzkum.

2.2.6.1 Sloučit, nebo nesloučit jednotlivé výskyty instancí

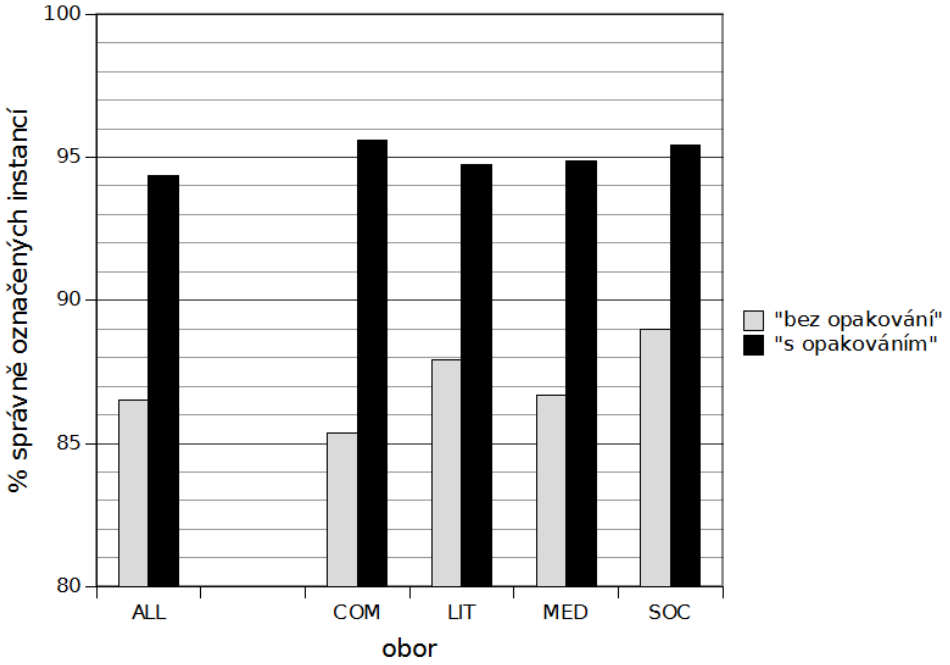
V data-miningovém nástroji Weka (viz kap. 2.3.3) je měřena úspěšnost metody J48graft (viz kap. 2.3.4.1) v označování termínů a netermínů v trénovacích datech „s opakováním“ a „bez opakování“. Následně se na stejných datech porovnávala také úspěšnost metody J48graft oproti srovnávací metodě ZeroR při použití dat „s opakováním“ a „bez opakování“. Textové pozice v trénovacích datech byly ve formě slovních tvarů, byly použity všechny vlastnosti kromě informace o zařazení do slovního druhu. Hranice mezi termíny a netermíny je v nástroji Weka defaultně nastavena na hodnotu 0,5[29].

V případě, že jednotlivé textové pozice jdou v tabulce za sebou stejně jako v textu, budeme je označovat jako data „s opakováním“ (jednotlivé instance se v datech můžou opakovat v různých řádcích, podle toho, jak se reálně vyskytují v textu), pokud jsou výskyty instancí sloučeny do jediného řádku (a pak řazeny např. abecedně), data zde nazýváme „bez opakování“.

V datech „bez opakování“ je daleko nižší počet řádků v tabulce, a to zhruba poloviční v případě, že jsou použity slovní tvary, a zhruba třetinový v případě lemmat (v datech „s opakováním“ je počet řádků u tvarů i lemmat stejný).

[29]Hranice mezi termíny a netermíny je sama předmětem zkoumání, viz kap. 3.3.1.

Obrázek 2.1: Úspěšnost (hodnota míry *accuracy*) metody J48graft na materiálu „bez opakování“ a „s opakováním“ (nelemmatizovaný materiál, bez morfologické anotace). Úspěšnost na materiálu „s opakováním“ je ve všech případech výrazně vyšší.



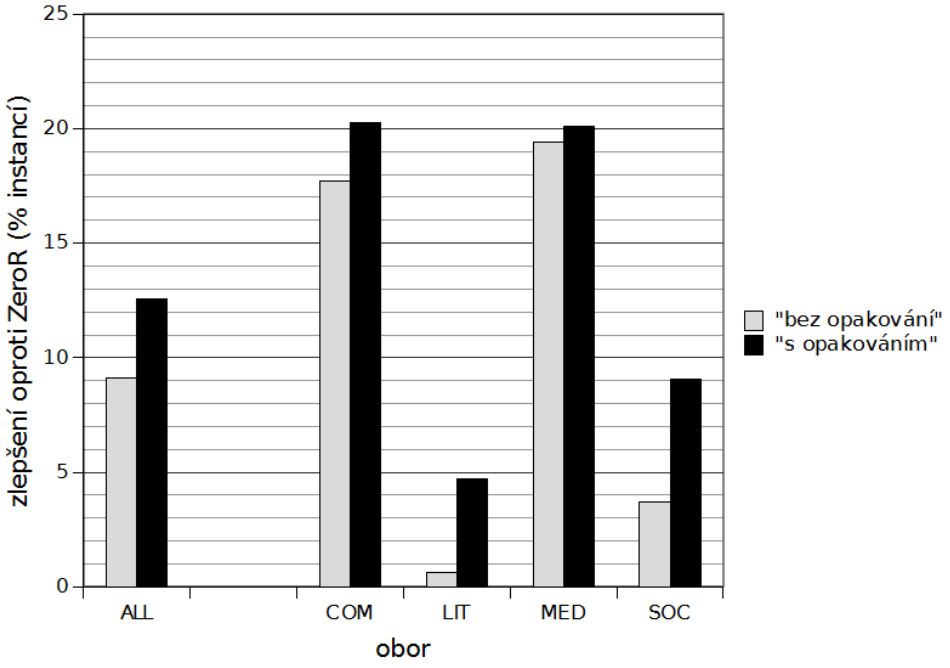
Z obrázku 2.1 vyplývá, že úspěšnost automatického označování termínů v trénovacích datech „s opakováním“ je ve všech oborech výrazně vyšší než v trénovacích datech „bez opakování“. Nezanedbatelný podíl na úspěšnosti má to, že některé jasné netermíny (např. interpunkce) se v textech velmi často opakují[30]. Nicméně opakování jednotlivých textových pozic zvyšuje úspěšnost i reálně: frekventovanější textové pozice, jako jsou interpunkce, gramatická slova, ale i časté termíny, napomáhají svou vysokou frekvencí učícímu procesu[31].

Obrázek 2.2 poskytuje doplňující informaci o tom, jaký je rozdíl mezi výsledky metody J48graft a srovnávací metody ZeroR (právě tento rozdíl mnohem přesněji ukazuje reálnou úspěšnost dané metody, viz kap. 2.4). I zde jsou výsledky poměrně přesvědčivé: trénovací data „s opakováním“ přinášejí ve všech oborech výraznější zlepšení oproti metodě ZeroR než data „bez opakování“ (větší rozdíly mezi obory nejsou dány podobou dat, ale povahou oborů, zvláště množstvím termínů v textech).

[30]Pokud je např. spojka *a* správně zařazena mezi netermíny jednou (v materiálu „bez opakování“), jistě to má na celkovou úspěšnost menší vliv, než když je takto správně zařazena opakovaně (v materiálu „s opakováním“).

[31]I v případě, že výsledná data rozdělená automaticky na termíny a netermíny vynásobíme frekvencí výskytů jednotlivých pozic v původním textu, je výsledná úspěšnost stále nižší než při použití dat „s opakováním“.

Obrázek 2.2: Obrázek ukazuje zlepšení výsledků metody J48graft oproti metodě ZeroR na materiálu „bez opakování“ a „s opakováním“ (nelematizovaný materiál, bez morfologické anotace). Zlepšení oproti ZeroR (tedy reálná úspěšnost dané metody) je ve všech oborech vyšší v případě materiálu ”„s opakováním“”(výrazněji u humanitních oborů).



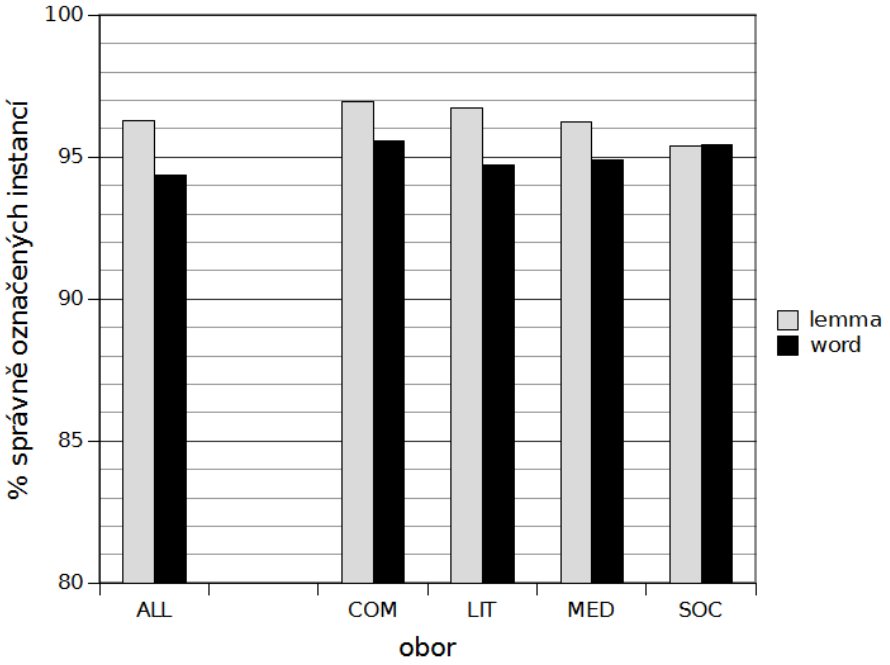
Použití dat „s opakováním“ vykazuje ve všech oborech lepší výsledky. Zároveň je použití jednotlivých textových pozic tak, jak jdou za sebou, v souladu s corpus-driven přístupem k datům (tj. co nejmenší zasahování do dat, práce se skutečným textem v původní podobě). Proto je vhodnější používat data „s opakováním“, tj. tak, jak jdou po sobě v reálném textu.

2.2.6.2 Instance ve formě lemmatu, nebo slovního tvaru

V nástroji Weka je srovnávána úspěšnost metody J48graft na materiálu lemmatizovaném a nelematizovaném (trénovací data). Doplňující informace poskytuje rozdíl úspěšnosti výsledků metody J48graft oproti srovnávací metodě ZeroR při použití dat lemmatizovaných a nelemmatizovaných. Textové pozice v trénovacích datech jsou ve formě „s opakováním“, byly použity všechny vlastnosti kromě informace o zařazení do slovního druhu. Hranice mezi termíny a netermíny je v nástroji Weka defaultně nastavena na hodnotu 0,5.

Jednotlivé instance v tabulce připravené pro data-miningové nástroje jsou tvořeny buď konkrétními slovními tvary použitými v textu, nebo lemmaty přiřazenými textovým pozicím v korpusu. Z hlediska přístupu řízeného daty (corpus-driven) by bylo vhodnější používat slovní tvary, tak jak se objevují v textu. Na druhou stranu je ale třeba dosáhnout vysoké míry úspěšnosti automatického vyhledávání, aby jakékoli závěry z vyvozované výsledků byly co nejspolehlivější.

Obrázek 2.3: Úspěšnost (hodnota míry *accuracy*) metody J48graft na materiálu lemmatizovaném a nelemmatizovaném (materiál „s opakováním“, bez morfologické anotace). Kromě oboru SOC (sociologie) je úspěšnost vždy vyšší na materiálu lemmatizovaném.



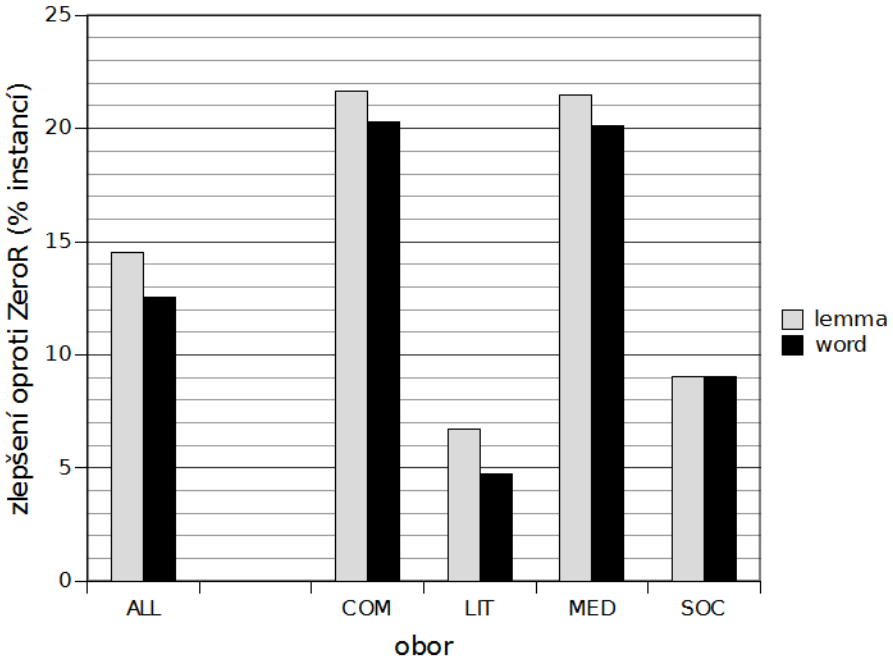
Proto je třeba zjistit, jak velký vliv má použití slovních tvarů oproti lemmatům na úspěšnost automatického vyhledávání termínů. Pokud by snížení úspěšnosti bylo natolik významné, že by tím byly zastíněny výhody corpus-driven přístupu, bylo by třeba v dalších experimentech používat lemmatizaci.

Z obrázku 2.3 je zřejmé, že užití lemmatizace zlepšuje výsledky data-miningové metody ve většině případů o jedno až dvě procenta (výjimkou je obor sociologie). Protože jde o rozdíl poměrně významný (viz kap. 2.4), je třeba posoudit, zda úspěšnost dosaženou na základě nelemmatizovaného textu můžeme ještě považovat za dostatečnou (k vyvozování závěrů o termínech). V tomto případě považujeme výhody přístupu řízeného korpusovými daty za převažující nad výhodami vysoké úspěšnosti vyhledávání termínů (už jen z čistě praktického hlediska jde o možnost využít jakýkoli, tedy i nelemmatizovaný text).

Obrázek 2.4 doplňuje informaci v obrázku 2.3 o srovnání rozdílu mezi metodou J48graft a metodou ZeroR. I zde jsou data s lemmatizací úspěšnější[32].

[32]Výjimkou je sociologie, kde je úspěšnost téměř vyrovnaná při použití slovních tvarů i lemmat.

Obrázek 2.4: Zlepšení výsledků metody J48graft oproti metodě ZeroR na materiálu lemmatizovaném a nelemmatizovaném (materiál „s opakováním“, bez morfologické anotace). Zlepšení oproti ZeroR (tedy reálná úspěšnost dané metody) je vyšší u materiálu lemmatizovaného (s výjimkou oboru SOC (sociologie)).



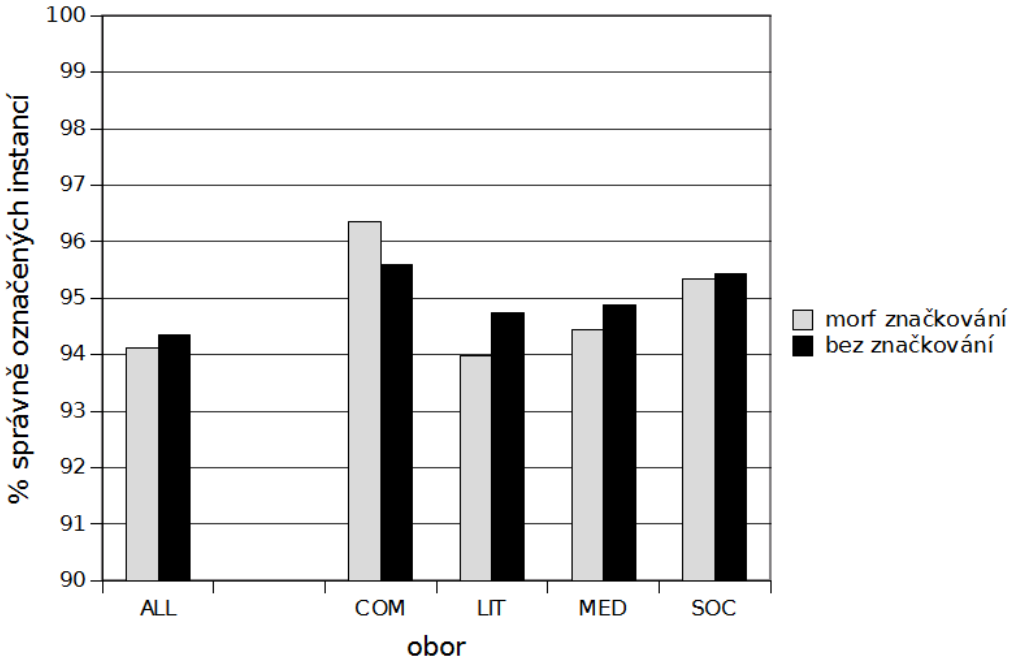
2.2.6.3 Využití morfologického značkování

V nástroji Weka je srovnávána úspěšnost dvou metod (J48graft a PART) na materiálu s určenými slovními druhy a bez nich (ostatní vlastnosti jsou v experimentu použity všechny). Materiálem jsou trénovací data, a to jak lemmatizovaná, tak i nelemmatizovaná. Jsou ve formě „s opakováním“. Hranice mezi termíny a netermíny je v nástroji Weka defaultně nastavena na hodnotu 0,5.

Jednou z vlastností, které jsou často využívány v metodách automatického vyhledávání termínů, je zařazení do slovního druhu (často jde pouze o rozlišení substantiv od nesubstantiv). Termíny totiž patří zdaleka nejčastěji ke slovnímu druhu substantiv, i když nezanedbatelně často i k adjektivům a slovesům, příp. i adverbiím; jen ve zvláštních případech i k jiným slovním druhům (k tomu viz kap. 1.2.2.2). I proto, že využívání morfologicky označkovaného materiálu je nevýhodnější z hlediska nároků na použité texty, je třeba podrobněji přezkoumat, jestli je zařazení do slovního druhu pro úspěšné vyhledání termínů skutečně důležité.

Obrázky 2.5 až 2.8 srovnávají úspěšnost vyhledávání termínů v materiálu morfologicky označkovaném a neoznačkovaném (tedy se slovními druhy a bez slovních druhů). Ze srovnání obrázků je zřejmé, že zařazení slovních druhů mezi atributy nehraje v úspěšnosti metod velkou roli, a to ani kladnou, ani zápornou. V různých kombinacích metod a dis-

Obrázek 2.5: Úspěšnost (*accuracy*) metody J48graft na materiálu morfologicky označkovaném a neoznačkovaném (určení slovních druhů) (nelemmatizovaný materiál, „s opakováním“). Úspěšnost je vyšší na materiálu bez označení slovních druhů, s výjimkou oboru COM (informatika).



ciplín je někdy úspěšnost na materiálu se slovními druhy lepší, někdy horší. Rozdíl přitom není nijak zásadní a většinou nepřekračuje jedno procento hodnoty míry *accuracy*. V tomto případě jednoznačně převažují výhody corpus-driven přístupu (a také např. stejná šance pro vyhledání termínů řazených mezi jiné slovní druhy, než jsou substantiva[33]).

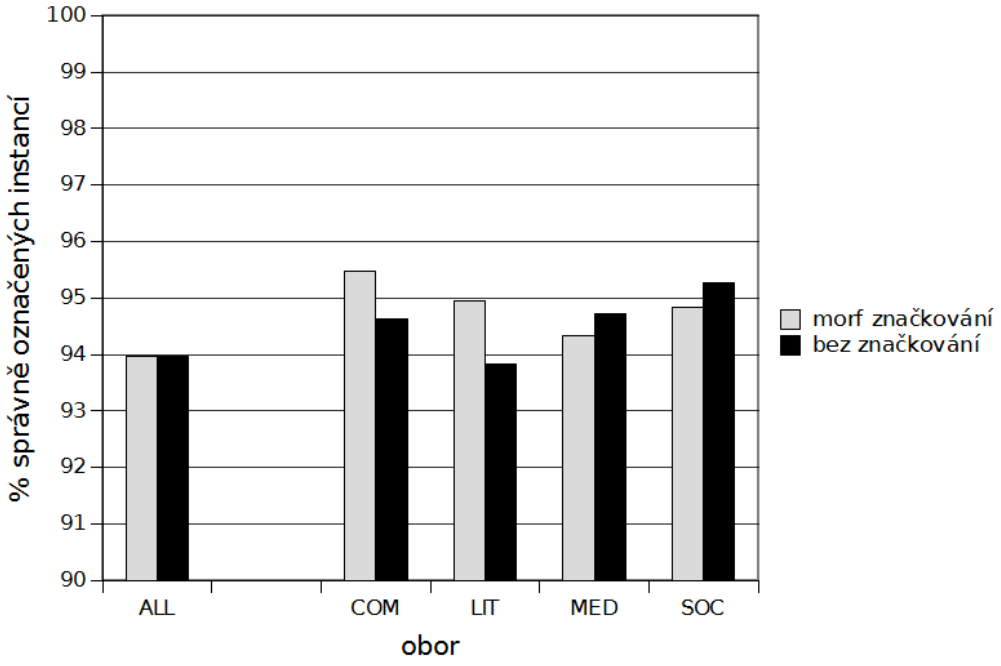
2.2.6.4 Shrnutí: nejvhodnější forma trénovacích dat

Experimenty byly zaměřeny na hledání co nejvhodnější podoby materiálu (trénovacích dat), na němž se budou provádět ostatní experimenty. Základem bylo vyvážení dvou aspektů: metoda vyhledávání termínů by měla mít vysokou úspěšnost a měla by být pokud možno co nejvíc řízena korpusovými daty, mj. proto, že tím se stává univerzálnější (z hlediska lingvistických teorií i z hlediska použitelných textů, více v kap.1.1.3).

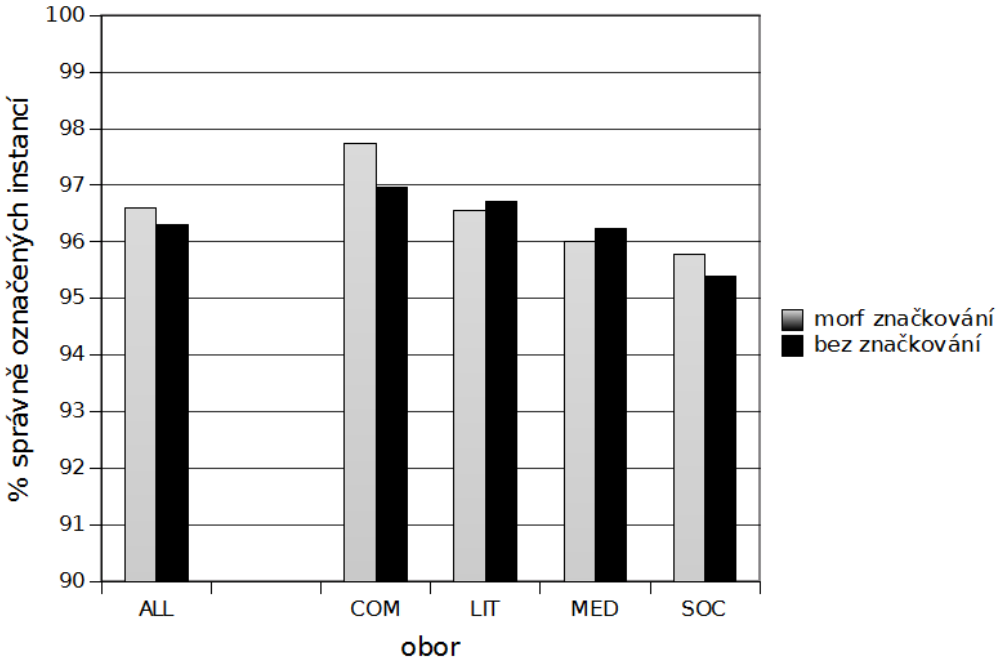
Bylo třeba zodpovědět tři otázky: 1. Sloučit, nebo nesloučit jednotlivé výskyty instancí (data „s opakováním“, nebo „bez opakování“)?, 2. Využít, nebo nevyužít lemmatizaci? a 3. Využít, nebo nevyužít morfologického značkování (slovní druh)?

[33]Takové termíny jsou sice okrajové, málo frekventované, neobvyklé, ale o to jsou zajímavější (viz kap. 1.2.2.2).

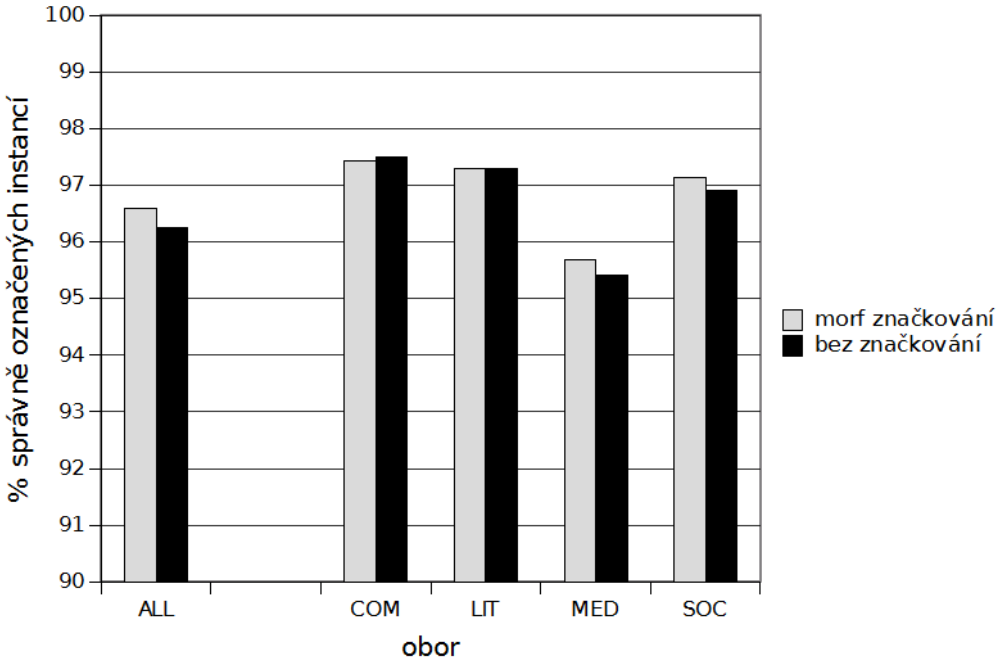
Obrázek 2.6: Úspěšnost (*accuracy*) metody PART na materiálu morfologicky označkovaném a neoznačkovaném (určení slovních druhů) (nelemmatizovaný materiál, „s opakováním“). Ne/výhodnost určení slovních druhů není prokazatelná.



Obrázek 2.7: Úspěšnost (*accuracy*) metody J48graft na materiálu morfologicky označkovaném a neoznačkovaném (určení slovních druhů) (lemmatizovaný materiál, „s opakováním“). Ne/výhodnost určení slovních druhů není prokazatelná.



Obrázek 2.8: Úspěšnost (*accuracy*) metody PART na materiálu morfologicky označkovaném a neoznačkovaném (určení slovních druhů) (lemmatizovaný materiál, „s opakováním“). Ne/výhodnost určení slovních druhů není prokazatelná.



1. Experimenty ukázaly, že slučování výskytů instancí je nevýhodné, protože nesloučená data („s opakováním“) poskytují vždy výrazně lepší výsledky. Co nejmenší zasahování do dat odpovídá corpus-driven přístupu, není tudíž třeba se rozhodovat: v dalších experimentech se budou používat **data ve formě „s opakováním“**.

2. Problematičtější je rozhodování, zda využívat, nebo nevyužívat lemmatizaci. Ta totiž ve většině případů zlepšuje výsledky oproti materiálu bez lemmatizace. Zlepšení je zdánlivě málo významné (většinou o jedno, maximálně dvě procenta); rozdíl mezi 95 a 97 procenty je ale ve skutečnosti velmi dobrým výsledkem (viz kap. 2.4). V následujících experimentech budou sice užívána **data bez lemmatizace**, ale pouze na základě rozhodnutí, že je v tomto případě dána přednost lingvistickému přístupu před úspěšností automatického vyhledávání termínů.

3. Využívání morfologického značkování, konkrétně zařazení do slovního druhu, vede u některých oborů k lepším a u některých k horším výsledkům[34]. Proto i zde je rozhodnutí jed-

[34]Zjištění, že morfologické značkování v určitých typech výzkumů nehraje zásadní roli, je zajímavé pro korpusovou lingvistiku. Podobně důležité je i zjištění o lemmatizaci, která výsledky automatického vyhledávání termínů zřetelně zlepšuje. Pro ucelenější přehled o celé problematice by ale bylo třeba věnovat se cíleně vlivu lingvistického značkování na různé typy výzkumů.

noduché: v příštích experimentech **není nutno využívat morfologického značkování** (což je výhodné i z hlediska hledání termínů v dalších textech).

2.3 Metoda

2.3.1 Nová metoda vyhledávání termínů TERMIT

Pro automatické vyhledávání termínů byla vyvinuta nová metoda TERMIT (Term Mining Tool). Ta je založena na používání data-miningových metod a nástrojů a je zaměřena primárně na získání nových poznatků o termínech a jejich vlastnostech, a až v druhé řadě na co nejvyšší úspěšnost při vyhledávání termínů.

Metoda TERMIT zahrnuje nalezení vhodné data-minigové metody z nástroje Weka (viz kap. 2.3.4.2), identifikaci nejdůležitějších kvantitativních vlastností termínů a sestavení kombinace menšího počtu vlastností efektivní při vyhledávání termínů (kap. 5) a postup při rozpoznávání jednoslovných a víceslovných termínů (viz níže).

Webové rozhraní metody TERMIT je k dispozici na internetových stránkách Českého národního korpusu (https://trnka.ff.cuni.cz/∼kovarikova/cgi-bin/termit.cgi). Pomocí aplikace je možné vyhledat termíny v jakémkoli textu a zároveň určit, k jakému oboru daný text pravděpodobně náleží.

2.3.2 Postup při automatické identifikaci termínů

Při automatické identifikaci termínů pomocí metody založené na data miningu je nutno podniknout celou řadu po sobě následujících kroků. Přehledné shrnutí jednotlivých kroků napomůže pochopení celého procesu. Ten je rozdělen do dvou částí, v první se identifikují jednoslovné termíny, a teprve na základě vyhledaných jednoslovných termínů se pokračuje v rozpoznávání termínů víceslovných:

A. Identifikace jednoslovných termínů v textech

1. příprava trénovacích dat: ruční přiřazení hodnot termín či netermín a automatické přiřazení hodnot jednotlivých vlastností (vztahujících se k jednoslovným termínům) všem textovým pozicím ve vybraných textech (kap. 2.2.3)

2. příprava testovacích dat: automatické přiřazení hodnot jednotlivých vlastností (vztahujících se k jednoslovným termínům) všem textovým pozicím ve vybraných textech (podle stejných principů jako v kap. 2.2.3)

3. natrénování vybraných metod z data-miningového nástroje na trénovacích datech

4. krosvalidace[35] na trénovacích datech ke zjištění úspěšnosti při identifikaci jednoslovných termínů (kap. 3.1.1)

5. seřazení vlastností podle důležitosti v procesu automatického vyhledávání termínů (kap. 5.1.1)

6. sestavení efektivní sady vlastností pro automatické vyhledávání jednoslovných termínů v testovacích datech (kap. 5.1.3)

7. stanovení (resp. ověření[36]) hranice mezi termíny a netermíny a identifikace jednoslovných termínů (kap. 3.3.1)

8. automatické přiřazení hodnoty terminologické platnosti jednotlivým textovým pozicím v testovacích datech (kap. 4.1.2), popř. i zpětně v datech trénovacích (kap. 4.3)

B. Identifikace víceslovných termínů v textech

1. příprava trénovacích dat: ruční přiřazení hodnot víceslovný termín či není víceslovný termín a automatické přiřazení hodnot jednotlivých vlastností (vztahujících se k víceslovným termínům) všem bigramům ve vybraných textech (kap. 2.2.3)

2. příprava testovacích dat: automatické přiřazení hodnot jednotlivých vlastností (vztahujících se k víceslovným termínům) všem bigramům ve vybraných textech (vč. informace o tom, zda jednotlivé části bigramu jsou jednoslovné termíny či netermíny) (kap. 2.2.3)

3. natrénování vybraných metod z data-miningového nástroje na trénovacích datech

[35]Krosvalidace je metoda zjišťování, do jaké míry bude metoda úspěšná v nezávislých (testovacích) datech. Množina trénovacích dat je rozdělena na podmnožiny. Jedna podmnožina slouží jako testovací množina, zbylé podmnožiny slouží jako trénovací množiny. Model se natrénuje na trénovací množině a pomocí testovací množiny se testuje přesnost a výkonnost tohoto modelu. Tento proces se několikrát opakuje, pokaždé s jinou podmnožinou tvořící trénovací a testovací množinu.

[36]Hranice mezi termíny a netermíny je v nástroji Weka defaultně nastavená na hodnotu 0,5.

4. krosvalidace (viz výše) na trénovacích datech ke zjištění úspěšnosti jednotlivých metod při identifikaci víceslovných termínů (kap. 3.2)
5. výběr nejdůležitějších vlastností (kap. 5.2.1)
6. sestavení efektivní sady vlastností pro automatické vyhledávání víceslovných termínů v testovacích datech (kap. 5.2.3)
7. stanovení (resp. ověření) hranice mezi termíny a netermíny a identifikace víceslovných termínů (kap. 3.3.2)
8. automatické přiřazení hodnoty terminologické platnosti jednotlivým bigramům v testovacích datech (kap. 4.4)

2.3.3 Data-miningový nástroj Weka

Pro automatické vyhledávání termínů a vyhodnocování důležitosti jednotlivých atributů byl použit především data-miningový nástroj Weka. Ten více či méně samostatně zpracovává trénovací data, tj. vybrané části akademických textů z korpusu SYN2010, kde jsou ručně označeny termíny jednoslovné a víceslovné a kde je automaticky jednotlivým textovým pozicím přiřazeno velké množství atributů.

Weka (Waikato environment for knowledge analysis) je data-miningový nástroj, který zahrnuje skupinu algoritmů pro analýzu dat a prediktivní modelování (Hall et al., 2009). Podporuje standardní data-miningové úlohy: klasifikaci, regresi, sdružování (clustering), hodnocení důležitosti atributů (feature ranking a feature selection) a jako pomocný nástroj také vizualizaci. Weka nabízí práci s velkým počtem metod, mezi kterými si uživatel může vybírat podle daného výzkumného úkolu - jednotlivé metody lze porovnat podle výsledků dosažených pro daný úkol. Weka poskytuje hodnocení výsledků pomocí krosvalidace nebo rozdělení dat na trénovací a testovací sadu, a to pro každou z nabízených metod.

2.3.4 Srovnání data-miningových metod

2.3.4.1 Seznam a popis použitých metod

Na základě předchozího výzkumu (Šrajerová, 2009a,b; Šrajerová et al., 2009), který se mimo jiné zabýval srovnáním data-miningových metod v různých data-miningových

nástrojích, bylo vybráno několik metod z nástroje Weka, jejichž úspěšnost byla nejvyšší, nebo které jsou vhodné pro účely srovnávání. Tyto metody se od sebe liší způsobem zpracovávání informací a lze je také různě využít (Witten, Frank, 2005). V následující kapitole se bude srovnávat jejich úspěšnost ve vyhledávání termínů v zadaných trénovacích datech a zároveň i jejich využitelnost pro další úkoly (feature ranking nebo další vyhledávání v libovolných datech).

ZeroR je nejjednodušší klasifikační metoda, která pouze označí všechny instance jako příslušníky majoritní třídy. V našem případě označí všechny textové pozice jako netermíny, protože ty v datech převládají. Její úspěšnost se tedy rovná procentu netermínů v daných datech. V oboru informatika (COM) obsazují netermíny zhruba 75 % textových pozic, úspěšnost metody ZeroR se tedy rovná cca 75 %. V oboru literatura (LIT) je ale netermínů asi 90 %, proto je úspěšnost metody ZeroR daleko vyšší. Metoda ZeroR neposkytuje žádné predikce o datech (není tedy použitelná pro vyhledávání termínů v textech), protože nepracuje s žádnými z poskytnutých atributů. Přesto je velmi důležitá, a to z toho důvodu, že slouží jako srovnávací základ pro hodnocení úspěšnosti ostatních metod. Všechny ostatní metody je třeba hodnotit na základě toho, jakého zlepšení dosáhly oproti této nejjednodušší metodě[37] (Witten, Frank, 2005).

J48graft je rozhodovací strom. Pro natrénování (a následné vyhledávání termínů) používá model ve formě stromu, který v jednotlivých větvích naznačuje, jaké má konkrétní rozhodnutí důsledky nebo výsledky. Jednoduše si takové rozvětvení stromu můžeme představit následovně: pokud je hodnota atributu X rovna určité hodnotě, pak je daná textová pozice termínem, pokud je menší, pak jde o netermín, pokud je větší, je třeba rozhodnout na základě atributu Y atd.

PART je metoda vytvářející pravidla na základě částečných rozhodovacích stromů. Jednoduchým příkladem pravidla je: pokud je hodnota atributu X rovna určité hodnotě a zároveň hodnota atributu Y je rovna určité hodnotě, pak je daná textová pozice termínem.

Bagging-PART (Weka) kombinuje rozhodnutí několika různých modelů, spojuje rozdílné výsledky do jediné predikce; buď jednotlivé metody hlasují, nebo se počítá průměr

[37] Pokud tedy metoda dosáhne v oboru COM 90% úspěšnosti, zlepšení bude značné (15 %) a použití metody je prodaný úkol vhodné. Pokud dosáhne 90% úspěšnosti v oboru LIT, nedojde v podstatě k žádnému zlepšení oproti metodě ZeroR, a tudíž metoda není pro daný úkol použitelná.

(v tomto případě se počítá průměr, protože výsledná hodnota pro každé slovo je numerická - získáme tedy škálu).

JRip také vytváří pravidla (na základě algoritmu nazvaného RIPPER), její fungování je tedy v principu podobné metodě PART. Mezi zkoumané metody je zařazena na základě velmi podobných výsledků, jako vykazuje metoda PART.

BayesNet patří mezi bayesovské metody, které jsou čistě statistické. Do tohoto srovnání metod byla zahrnuta jen pro doplnění, protože úspěšnost bayesovských metod je v porovnání se stromy nebo pravidly v případě tohoto výzkumu mnohem nižší (nebudou tedy v dalších výzkumech využívány).

Lineární regrese je nejjednodušší model, tj. proložení dat přímkou. Z principu nemůže zachycovat vztahy složitější než lineární, na rozdíl od např. stromů tedy nedokáže využít jeden atribut víckrát, proto má vždy vyšší úspěšnost s více atributy. Výhodou metody je, že na rozdíl od většiny použitých metod umí vytvořit poměrně jednoduchou rovnici pro automatické vyhledávání termínů (ovšem kvůli svým omezením méně úspěšnou ve srovnání s vyhledáváním pomocí jiných metod).

2.3.4.2 Srovnání úspěšnosti dostupných metod

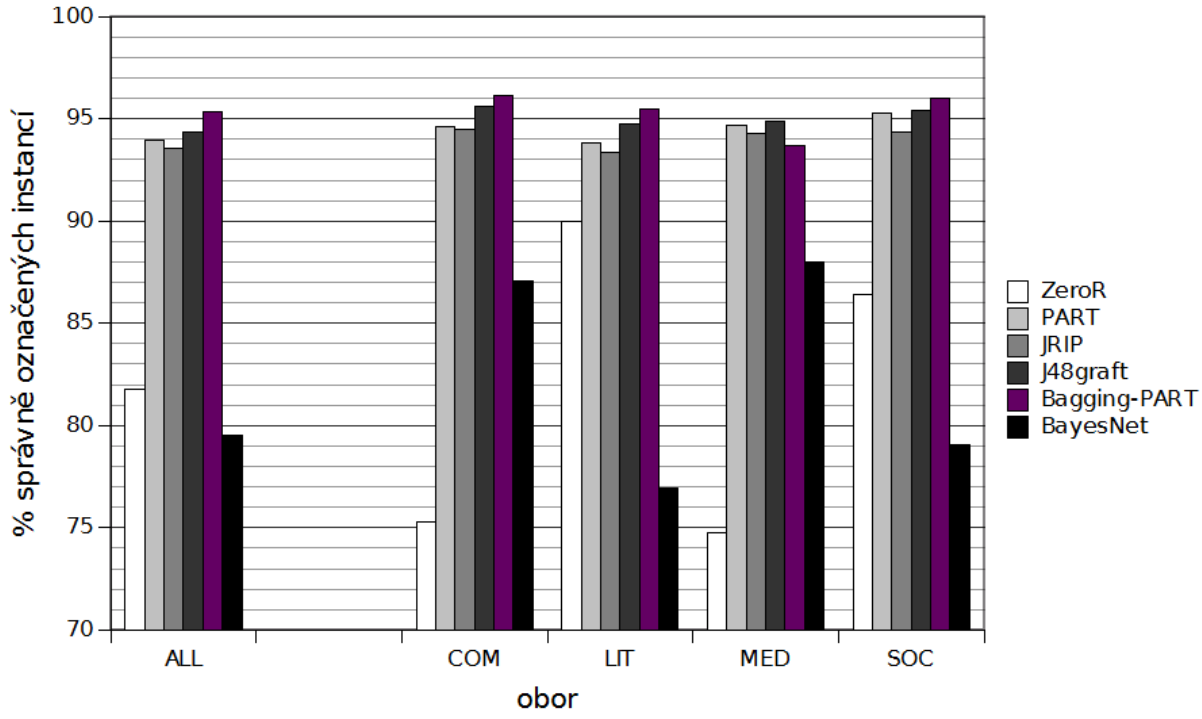
V data-miningovém nástroji Weka je srovnávána úspěšnost metod ZeroR, J48graft, PART, JRIP, Bagging-PART a BayesNet na trénovacích datech. Materiál je ve formě nelemmatizované, „s opakováním“ a bez rozlišení slovních druhů. V experimentu byly použity všechny vlastnosti (kromě informace o zařazení do slovního druhu). Hranice mezi termíny a netermíny je v nástroji Weka defaultně nastavena na hodnotu 0,5.

Jedním z cílů této práce je ověřit, zda a s jakou úspěšností jsou data-miningové metody schopné automaticky vyhledat termíny v předložených textech. Na základě výsledků metod s vysokou úspěšností vyhledávání totiž pak lze vyvozovat spolehlivější závěry o charakteristikách termínů[38].

Obrázek 2.9 porovnává úspěšnost vybraných metod na materiálu bez lemmatizace, bez rozlišení slovních druhů a „s opakováním“ (viz kap. 2.2.6). Výsledky metody ZeroR slouží pouze pro srovnání. Metoda BayesNet je zařazena jen pro úplnost (aby mezi metodami byla alespoň jedna čistě statistická), ale její úspěšnost je výrazně nižší než u ostatních metod (dále tedy nebude používána). Nejlepší výsledky vykazuje metoda Bagging-PART. Ostatní tři metody jsou zhruba stejně úspěšné (a neliší se nijak zásadně ani při experimentech s odlišně nastavenými, např. lemmatizovanými daty).

[38] Jednotlivé výsledky se liší nejen podle toho, jaké konkrétní metody jsou vybrány pro daný experiment,

Obrázek 2.9: Srovnání úspěšnosti vybraných metod z nástroje Weka. S výjimkou oboru SOC je nejúspěšnější metoda Bagging-PART, nejméně úspěšná je metoda BayesNet. Metoda ZeroR je použita pro srovnání. Materiál: nelemmatizovaný, „s opakováním“, bez určení slovních druhů.



2.3.4.3 Shrnutí: úspěšnost metod

Srovnání dostupných metod ukázalo, že nejvyšší úspěšnost má jednoznačně metoda Bagging-PART, proto je vhodné ji používat v dalších experimentech jako součást metody vyhledávání termínů TERMIT. Nejhorší výsledky poskytuje podle očekávání statistická metoda BayesNet (ta v dalších experimentech už nebude užívána vůbec). Zbývající tři zkoumané metody J48graft, PART a JRip jsou zhruba stejně úspěšné ve všech oborech i ve všech nastaveních dat, v některých experimentech je lze využít vedle nejúspěšnější metody Bagging-PART.

2.3.5 Přehled experimentů

Pomocí data-miningového nástroje Weka byla provedena řada experimentů (kompletní seznam a podrobný popis experimentů je v kapitolách 4 a 5), jejichž úkolem je odpovědět na následující otázky:

- Jaká je úspěšnost vyhledávání jednoslovných i víceslovných termínů v jednotlivých disciplínách (jak je data mining vhodný k automatickému vyhledávání termínů)?
- Jak odpovídají automaticky vyhledané termíny ručnímu vyhodnocení? (kap. 3.1.1 a 3.2)
- Které atributy se největší měrou podílejí na automatickém vyhledávání jednoslovných nebo víceslovných termínů, které jsou tedy nejdůležitějšími vlastnostmi termínů? (kap. 5.1.1 a 5.2.1)
- Existuje vhodná kombinace malého počtu atributů, pomocí níž vyhledáme většinu jednoslovných nebo víceslovných termínů v textech? (kap. 5.1.3 a 5.2.3)
- Jaká je terminologická platnost automaticky vyhledaných termínů a netermínů? Funguje v terminologii princip škály? (kap. 3.3.1 a 3.3.2)
- Jaké množství jednoslovných termínů je v textech a ve slovní zásobě různých akademických disciplín? (kap. 4.1.2)
- Jaké množství víceslovných termínů je v textech a ve slovní zásobě různých akademických disciplín? (kap. 4.4)

ale i podle nastavení trénovacích dat (viz kap. 2.2.6) nebo podle příslušnosti k oboru (viz kap. 3.1.1).

- Lze na základě automaticky vyhledaných termínů zjistit rozdíly nebo podobnosti mezi disciplínami? (kap. 4.2)

2.4 Vyhodnocování úspěšnosti automatického vyhledávání

V následujících kapitolách se vyhodnocují a interpretují výsledky různých experimentů a srovnávají se data-miningové metody dostupné v nástroji Weka (viz výše). Ve většině případů se hodnotí **úspěšnost** určité metody. Význam slova úspěšnost je poměrně obecný, stejně tak by se dalo mluvit o tom, jak „dobře“ daná metoda označuje termíny a/nebo netermíny. Proto je třeba stanovit, co se úspěšností v rámci daného výzkumu míní.

V případě tohoto výzkumu se výsledky jednotlivých experimentů hodnotí standardními evaluačními statistickými mírami, jako je *precision, recall, F-measure* a především *accuracy*. Ve většině případů se zde slovem **úspěšnost** míní právě hodnota míry ***accuracy***, která sleduje správné zařazení všech textových pozic mezi termíny nebo netermíny; zabývá se tedy jak členy příznakovými, tak nepříznakovými. Ostatní evaluační míry berou v úvahu pouze členy příznakové - termíny, vyhodnocují tedy daleko menší část zkoumaných textových pozic (i proto je jejich hodnota výrazně nižší, než je tomu u hodnoty *accuracy*). Pokud je proces automatického vyhledávání termínů hodnocen jinou evaluační mírou než *accuracy*, je to vždy uvedeno.

Výhody a nevýhody jednotlivých měr jsou spolu se způsobem výpočtu přehledně zpracovány v publikaci Foundation of Statistical NLP (Manning, 2000)[39].

Vyhledávání termínů je tzv. binární klasifikace, při které se všechny zkoumané textové pozice zařazují do jedné ze dvou kategorií (termín a netermín). Výsledek je pak hodnocen podle toho, kolik skutečných termínů bylo správně označeno za termín (TP, true positive), kolik netermínů bylo správně začleněno do kategorie netermín (TN, true negative), kolik termínů bylo nesprávně zařazeno mezi netermíny (FN, false negative) a kolik netermínů bylo nesprávně označeno za termíny (FP, false positive).

Accuracy je míra posuzující podíl správných výsledků v populaci, tedy podíl správně označených textových pozic (ať už jde o termíny, nebo netermíny). 100% *accuracy* znamená, že naměřené hodnoty jsou přesně stejné jako skutečné hodnoty, tedy že všechny

[39]Vzorce i další údaje uvedené v této kapitole jsou z této knihy přebírány (Manning, 2000, s. 267-271).

termíny i všechny netermíny byly správně označeny. Rovnice pro výpočet *accuracy* je následující:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision je podíl relevantních instancí, které byly vyhledány, z celkového počtu vyhledaných instancí. V rámci této práce je to podíl správně vyhledaných termínů ku celkovému počtu automaticky označených termínů. 100% *precision* znamená, že všechny textové pozice označené jako termíny jsou skutečně termíny (ale je možné, že nebyly vyhledány veškeré termíny obsažené v textu). Rovnice pro výpočet míry *precision* je následující:

$$Precision = \frac{TP}{TP + FP}$$

Recall je podíl vyhledaných instancí, které jsou relevantní, tedy v tomto případě podíl správně vyhledaných termínů ku počtu všech termínů obsažených mezi textovými pozicemi. 100% *recall* znamená, že byly vyhledány a správně označeny všechny termíny, které se v rámci textových pozic vyskytly (ale je možné, že za termíny byly označeny i některé netermíny). Rovnice pro výpočet míry *recall*:

$$Recall = \frac{TP}{TP + FN}$$

F-measure kombinuje *precision* a *recall* v jediné míře. Ta je definována takto (pokud stanovíme, že *precision* a *recall* mají rovnocennou váhu):

$$F = \frac{2 \cdot Precision \cdot Recall}{(Recall + Precision)}$$

Při používání evaluačních měr je nutné si uvědomit, že rozdíl jednoho nebo dvou procent může mít velmi rozdílnou váhu. Rozdíl mezi 60 % a 61 % je z hlediska automatického vyhledávání termínů velmi malý, ať už jde o jakoukoli ze statistických měr - zlepšení oproti původnímu stavu je totiž jen 2,5 %. Naproti tomu rozdíl mezi 90 % a 91 % je výraznější (zlepšení o 10 %), a rozdíl mezi 96 % a 97 % je ještě zásadnější (zlepšení o 25 %).

Další poznámka vztahující se k úspěšnosti se týká porovnávání výsledků zkoumané me-

tody se **srovnávací metodou ZeroR** (Witten, Frank, 2005, s. 409)[40]. Jde o jednoduchou metodu, jejíž úspěšnost zcela závisí na podílu termínů a netermínů (nebo jiných typů) ve zkoumaném materiálu. ZeroR přiřadí všechny textové pozice do majoritní kategorie, tedy v našem případě do kategorie netermín, a tedy hodnota *accuracy* se rovná procentu netermínů v daném textu. S touto metodou je nutno porovnat výsledky každé jiné použité metody - úspěšnost metody se totiž neodvíjí jen od dosažených procent správně zařazených textových pozic, důležitý je i rozdíl mezi danou metodou a srovnávací metodou ZeroR (čím větší je rozdíl mezi ZeroR a další metodou, tím je tato metoda pro daná data úspěšnější).

Hodnocení úspěšnosti automatického zpracování dat je vždy relativní. Neexistuje žádná obecně platná mez, která by oddělovala úspěšné a neúspěšné metody. Záleží vždy na materiálu, který je zkoumán, v našem případě například na oboru, z něj pochází daný text, i na konkrétní úloze (vyhledávání jednoslovných a víceslovných termínů). Jediné srovnání, které je k dispozici, jsou výsledky metod zaměřené na podobný úkol a prováděné na podobném materiálu. I v takových případech je ale třeba počítat s tím, že úspěšnost se bude lišit na základě jiných vlivů, než je sama kvalita metody.

Jedním z příkladů vyhodnocování úspěšnosti při automatickém vyhledávání termínů je článek srovnávající pět metod automatického vyhledávání jednoslovných termínů (Marín, 2014). Nejvyšší dosažená hodnota míry *precision* při identifikaci termínů ve specializovaném korpusu právních textů je 73,5 % (Marín, 2014, s. 83) - to je tedy hodnota, ke které lze aspoň relativně vztahovat hodnoty *precision* dosažené ve předkládaném výzkumu (viz např. kap. 3.3).

[40] Více o metodě ZeroR v kap. 2.3.4.1.

3 Úspěšnost metody TERMIT při hledání termínů

Výzkum termínů v reálných textech je založen na dvou hypotézách, z nichž první se soustředí na možnosti automatického rozpoznávání termínů v textech. Metoda automatického vyhledávání termínů TERMIT založená na data-miningu dokáže identifikovat jednoslovné i víceslovné termíny s poměrně vysokou úspěšností. Právě hodnocení úspěšnosti vyhledávání termínů či zařazování textových pozic (zvláště slov) mezi termíny a netermíny je předmětem následující kapitoly.

Kapitola je rozdělena na dvě části, z nichž první je věnovaná různým pohledům na úspěšnost metody TERMIT u jednoslovných a víceslovných termínů, a to jak v datech trénovacích, tak i testovacích.

Druhá podkapitola se soustředí na stanovování hranice mezi termíny a netermíny na základě hodnoty terminologické platnosti automaticky přidělené jednotlivým textovým pozicím (jednoslovné termíny) či bigramům (víceslovné termíny). Tato hranice ovlivňuje právě úspěšnost identifikace termínů a podle různých výzkumných cílů je nutno ji přehodnocovat.

3.1 Úspěšnost při vyhledávání jednoslovných termínů

Úspěšnost automatického vyhledávání termínů lze hodnotit různým způsobem. Nejjednodušší je zjistit, kolik z celkového počtu zkoumaných textových pozic bylo správně zařazeno buď mezi termíny, nebo netermíny. To vyjadřuje evaluační míra *accuracy* (viz kap. 2.4). Jiným hlediskem je stejné zhodnocení správného zařazení textových pozic mezi termíny či netermíny, ovšem ve srovnání s výsledky metody ZeroR, která je nejjednodušší metodou určenou pro porovnávání různých datových sad (viz tamtéž). To nám umožní zjistit, jaký je prostor pro zlepšování úspěšnosti metody a jak ho bylo využito. V neposlední řadě je možné soustředit se pouze na samotné termíny, tedy na počet správně vyhledaných termínů, na počet netermínů zařazených mezi termíny a naopak termínů zařazených nesprávně mezi netermíny. K tomu lze využít evaluační míry *precision, recall* a *F-measure* (viz tamtéž).

3.1.1 Automatické vyhledávání v trénovacích datech

Parametry experimentu 1: Úspěšnost automatického vyhledání jednoslovných termínů v trénovacích datech. V experimentu je srovnávána úspěšnost metod Bagging-PART, J48graft a ZeroR v jednotlivých oborech v trénovacích datech (COM, LIT, MED, SOC; 8 tisíc textových pozic). Materiál je ve formě nelemmatizované, s opakováním a bez rozlišení slovních druhů. V případě metody J48graft byly použity všechny vlastnosti přiřazované jednoslovným termínům, v případě metody Bagging-PART pouze 4 vybrané vlastnosti (kap. 5.1.3). Na tomto základě byly vypočítány evaluační míry *accuracy, precision, recall* a *F-measure* (kap. 2.4). Hranice mezi termíny a netermíny je v nástroji Weka defaultně nastavena na hodnotu 0,5 (viz kap. 3.3).

Celková úspěšnost zařazování textových pozic mezi termíny a netermíny v trénovacích datech je velmi vysoká a ve všech oborech zhruba vyrovnaná. Pomocí metody J48graft je správně zařazeno necelých 95 % textových pozic, pomocí metody Bagging-PART se počet správně označených pozic blíží až 96 % (viz 3.1).

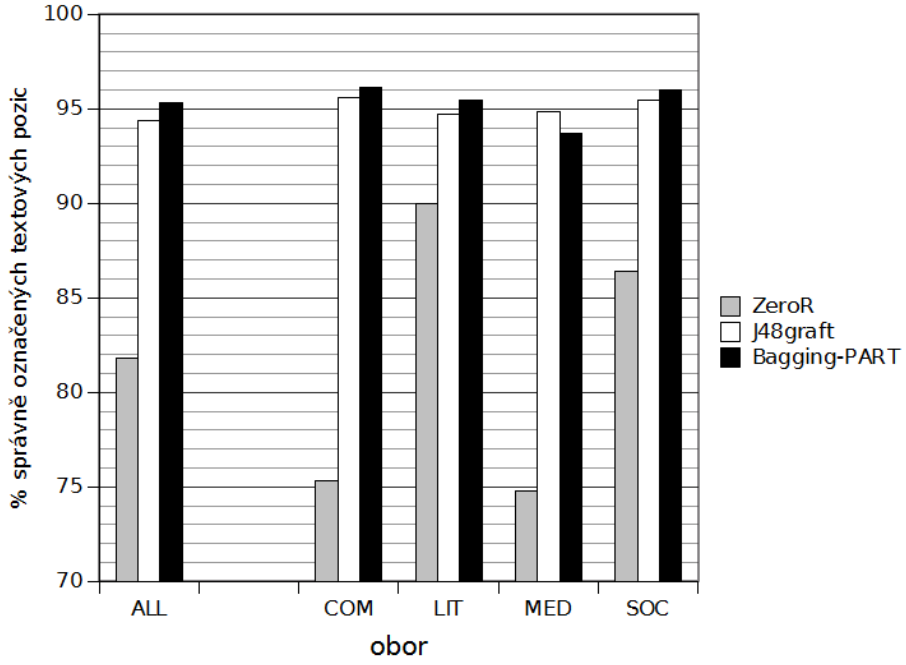
Srovnání těchto výsledků s výsledky metody ZeroR ve stejném obrázku přináší odlišný pohled na jednotlivé disciplíny. Zařazování textových pozic mezi termíny a netermíny je z tohoto pohledu výrazně úspěšnější v oborech COM a MED oproti humanitním oborům SOC a především LIT. Ač je tedy celková úspěšnost zařazování textových pozic o něco vyšší v oboru LIT než v oboru MED (v případě metody Bagging-PART), vyšší rozdíl mezi úspěšností a hodnotou ZeroR ukazuje na vyšší efektivitu vyhledávání termínů a netermínů v oboru MED (rozdíl 5 % oproti 20 %). Důvodem je odlišný prostor pro zlepšování úspěšnosti metody, založený na počtu termínů a netermínů v datech jednotlivých oborů[1].

Ještě jinak lze hodnotit úspěšnost metody, pokud se zaměříme pouze na samotné termíny. Obrázek 3.2 sleduje evaluační míry *precision* a *recall* při vyhledávání jednoslovných termínů v trénovacích datech za pomoci metody Bagging-PART. Největší procento správně označených termínů (precision: 95 %), stejně jako největší procento nalezených termínů (recall: 89 %) je v oboru COM, který byl z hlediska celkové úspěšnosti vyrovnaný s oborem SOC a z hlediska zlepšení oproti ZeroR s oborem MED.

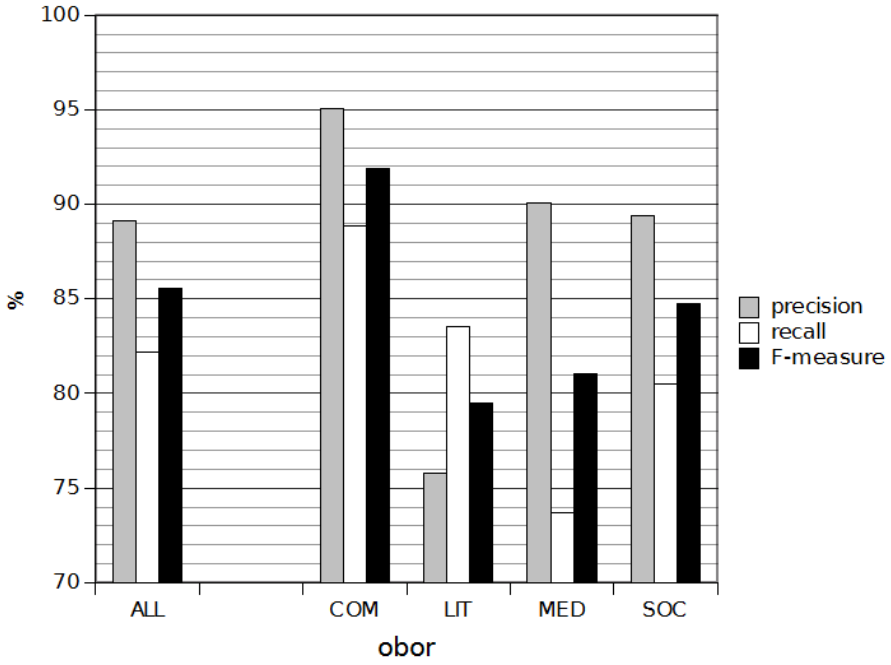
Na základě této analýzy lze tedy tvrdit, že největší úspěšnosti (v obecnějším slova smyslu) bylo dosaženo právě v oboru COM (informatika), protože až 96 % textových pozic bylo správně zařazeno mezi termíny a netermíny, rozdíl oproti ZeroR dosahuje více než 20 %, 95 % textových pozic označených za termíny jsou skutečně termíny a celkově bylo vyhledáno 89 % termínů obsažených ve zkoumaných textech oboru.

[1]Menší počet termínů v některých oborech má také negativní vliv na natrénování data-miningové metody, málo informací totiž ztěžuje proces učení.

Obrázek 3.1: Srovnání jednotlivých oborů na základě úspěšnosti (*accuracy*) automatického zařazování textových pozic mezi termíny a netermíny v trénovacích datech. Použité metody jsou J48graft a Bagging-PART. Pro srovnání jsou zařazeny i výsledky metody ZeroR. Nejvyšší úspěšnosti i největšího rozdílu oproti metodě ZeroR bylo dosaženo v oboru COM (informatika).



Obrázek 3.2: Srovnání úspěšnosti vyhledávání termínů v trénovacích datech na základě evaluačních měr *precision* a *recall*. Použitou metodou je Bagging-PART. Nejúspěšnější z tohoto pohledu je taktéž vyhledávání v oboru COM (informatika) (viz předcházející obrázek).



3.1.2 Automatické vyhledávání v rozsáhlých testovacích datech

Parametry experimentu 2: Srovnání úspěšnosti automatického vyhledání jednoslovných termínů v testovacích datech. V experimentu 2 jsou za pomoci metody Bagging-PART automaticky označeny termíny a netermíny v testovacích datech (37 oborů z korpusu SYN2010 s různým počtem textových pozic v každém oboru). Pro zjištění úspěšnosti byly v náhodně vybraných sto po sobě jdoucích pozicích zjištěny rozdíly mezi automatickým a ručním označením. Na tomto základě byly vypočítány evaluační míry *accuracy, precision, recall* a *F-measure* (kap. 2.4). Materiál je v podobě nelemmatizované, s opakováním a bez slovních druhů. V experimentu byly použity vlastnosti RFQdiscRFQcompar, ARF, RDist, a SDRD (viz kap. 5.1.3). Hranice mezi termíny a netermíny je v nástroji Weka defaultně nastavena na hodnotu 0,5.

Pomocí metody Bagging-PART (viz kap. 2.3.4.2) byly automaticky zpracovány všechny dostupné texty ve 37 akademických oborech z korpusu SYN2010 (viz tab. 2.1). Pro každou textovou pozici byla vypočtena hodnota terminologické platnosti na škále mezi hodnotami 0 a 1 (viz kap. 1.3); mezi termíny byly zařazeny ty textové pozice, které dosáhly hodnoty terminologické platnosti 0,5 a vyšší[2].

Zjišťování úspěšnosti automatického označování termínů probíhalo ručním ověřováním správnosti označení termínů a netermínů. Z každého oboru bylo náhodně vybráno sto za sebou jdoucích textových pozic a ručně byla ověřena správnost jejich automatického zařazení[3]. Pro všechny obory tak byla zjištěna úspěšnost automatického vyhledávání termínů, a to na základě evaluačních měr *accuracy, F-measure, precision* a *recall* (viz kap. 2.4).

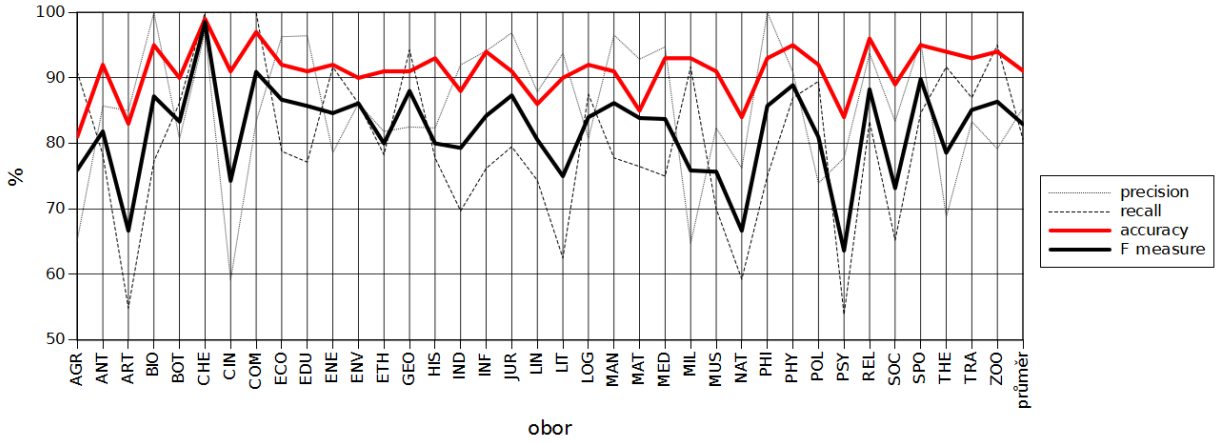
Výsledky shrnuté v obrázku 3.3 ukazují, že automatické vyhledávání termínů napříč množstvím různých oborů je poměrně úspěšné, zhruba 91 % ručně kontrolovaných textových pozic bylo správně zařazeno mezi termíny nebo netermíny. Průměrné procento termínů mezi všemi textovými pozicemi automaticky označenými jako termíny je 85 % (evaluační míra *precision*), průměrně 81 % všech termínů obsažených v ručně kontrolovaných datech bylo úspěšně vyhledáno (*recall*). Srovnání s metodou ZeroR (obrázek 3.4) ukazuje, že úspěšnost zařazování mezi termíny a netermíny (*accuracy*) není výrazně závislá na počtu termínů v oboru, který je vyjadřován právě hodnotou ZeroR (viz kap. 2.4).

Automatické vyhledání termínů v textech velkého počtu akademických disciplín umožňuje srovnání oborů z různých úhlů, především právě z pohledu úspěšnosti. Dále lze zjistit, kolik procent textu v různých oborech obsazují termíny a kolik procent ze všech užitých lemmat v určitém oboru tvoří termíny, případně zda se tyto údaje liší v různých typech oborů

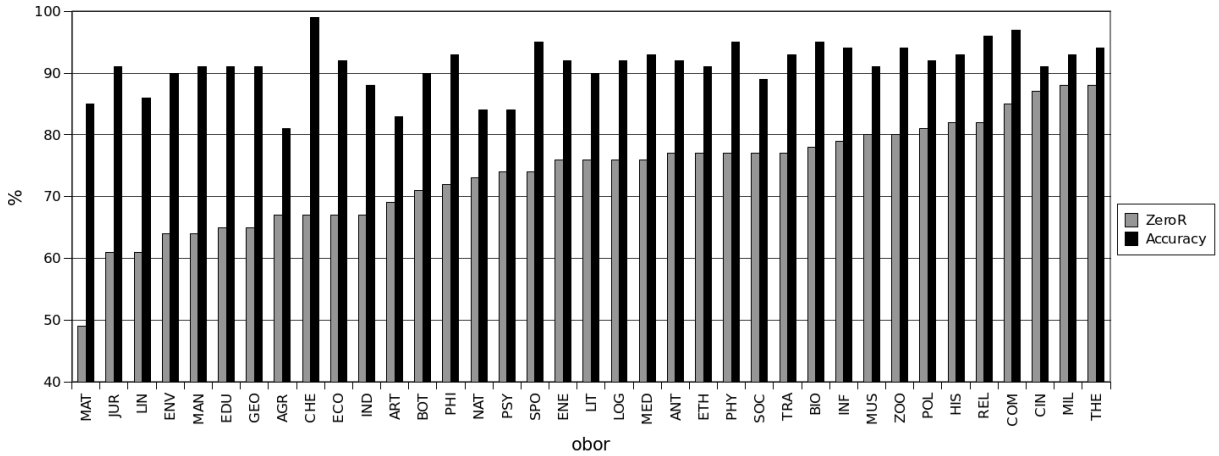
[2]Více o hranici mezi termíny a netermíny v kap. 3.3.1.

[3]Takto pojaté posouzení přesnosti vyhledávání v jednotlivých oborech má svá omezení, především se vztahuje pouze k vybraným textovým pozicím, přičemž tento náhodně vybraný krátký úsek textu může být pro danou disciplínu netypický. Daleko víc se už lze spolehnout na průměrnou hodnotu ze všech oborů, protože je založena na daleko vyšším počtu náhodně vybraných textových pozic (100 textových pozic z každého z 37 oborů).

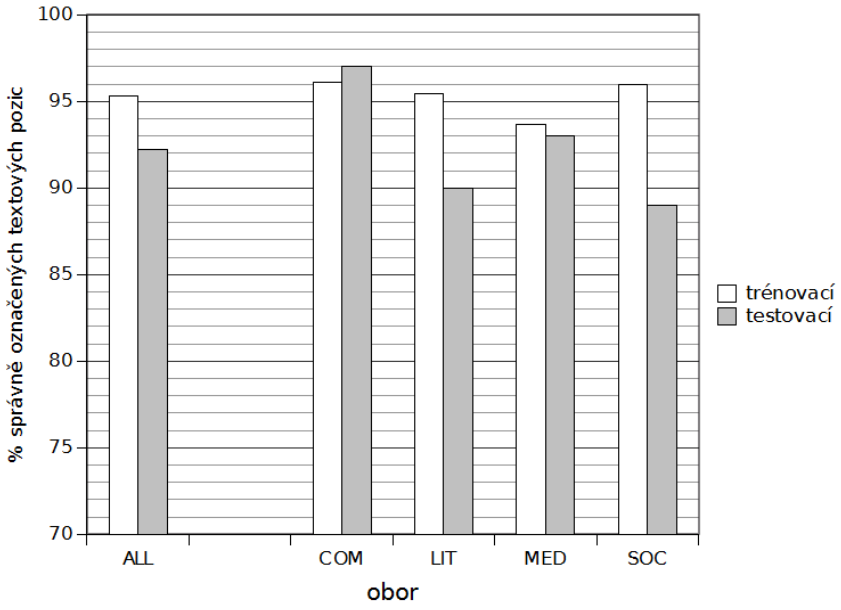
Obrázek 3.3: Úspěšnost automatického vyhledávání termínů ve 37 oborech dostupných v SYN2010. Průměrná úspěšnost zařazování textových pozic mezi termíny a netermíny (accuracy) je 91,1 %, průměrné hodnoty měr vyjadřujících správnost zařazování termínů jsou: precision = 85,4 %, recall = 80,6 % a F-measure = 82,9 %.



Obrázek 3.4: Srovnání úspěšnosti metody PART-Bagging (*accuracy*) a metody ZeroR (jejíž výsledek se rovná počtu ručně označených netermínů ve stejných 100 textových pozicích) ve 37 oborech dostupných v SYN2010. Ukazuje se, že úspěšnost označování termínů a netermínů není výrazně závislá na celkovém počtu termínů v textech oboru. Výsledky jsou tříděny podle hodnoty ZeroR.



Obrázek 3.5: Srovnání úspěšnosti na trénovacích a testovacích datech ze čtyř oborů.



(humanitní, přírodovědné) (viz tab. 4.5 v kap. 4.1.2).

Obrázek 3.5 porovnává výsledky automatické identifikace termínů v trénovacích a testovacích datech čtyř vybraných oborů: COM, LIT, MED a SOC. Vyšší úspěšnost v trénovacích datech (kromě oboru COM) je obvyklá a je dána tím, že metoda je natrénována právě na textových pozicích obsažených v trénovacích datech.

3.2 Úspěšnost vyhledávání víceslovných termínů

Automatické vyhledávání víceslovných termínů závisí na identifikaci termínů jednoslovných, z kterých jsou obvykle víceslovné termíny (alespoň z části) tvořeny. Přítomnost alespoň jednoho jednoslovného termínu je velmi důležitý atribut víceslovných termínů (viz kap. 2.3).

Stejně jako u jednoslovných termínů, i u těch víceslovných lze hodnotit úspěšnost automatického vyhledávání termínů různými způsoby. Prvním způsobem je zjišťování počtu textových pozic (zvláště slov) správně zařazených mezi termíny nebo netermíny (*accuracy*), druhým je srovnání tohoto údaje s výsledky ZeroR, které umožňuje zjistit, jaký je prostor pro zlepšování dané metody. Třetí možností je zaměřit se na samotné termíny, konkrétně na počet termínů nesprávně zařazených mezi netermíny (tedy nevyhledaných; *recall*) a na počet skutečných termínů mezi všemi automaticky označenými termíny (*precision*) (více

k vyhodnocování výsledků kap. 2.4).

3.2.1 Automatické vyhledávání v trénovacích datech

Parametry experimentu 3: Srovnání úspěšnosti automatického vyhledání víceslovných termínů v trénovacích datech. V experimentu 3 je srovnávána úspěšnost označování víceslovných termínů pomocí metody Bagging-PART v jednotlivých oborech v trénovacích datech. Materiál je ve formě bigramů, bez lemmatizace, s opakováním a bez rozlišení slovních druhů. V experimentu byly použity všechny vlastnosti určené pro vyhledávání víceslovných termínů (označené jako MWT). Hranice mezi termíny a netermíny je v nástroji Weka defaultně nastavena na hodnotu 0,5.

Úspěšnost (*accuracy*) metody Bagging-PART při automatickém zařazování bigramů mezi víceslovné termíny a netermíny je v trénovacích datech velmi vysoká (kolem 97 %). Ve skutečnosti je to ale dáno tím, že víceslovných termínů je v trénovacích datech poměrně malé množství, a tedy už jen pouhé zařazení všech textových pozic mezi netermíny (což odpovídá metodě ZeroR, viz kap. 2.4) vykazuje výsledky okolo 91 % - prostor pro zlepšování výsledků metody je tedy poměrně malý.

Obrázek 3.6 nabízí přehled úspěšnosti automatického vyhledávání víceslovných termínů v trénovacích datech 4 oborů a zároveň přehled rozdílů mezi výsledky metody Bagging-PART a ZeroR. Na rozdíl od vyhledávání jednoslovných termínů se úspěšnost zařazování mezi termíny či netermíny liší u jednotlivých oborů až o 2 % (96 % v COM a SOC vs. 98 % v MED).

Při srovnání s výsledky ZeroR zjišťujeme, že nejúspěšnější z tohoto pohledu je vyhledávání víceslovných termínů v oboru SOC, kde je rozdíl mezi výsledky metody Bagging-PART a ZeroR nejmarkantnější, nejméně úspěšné v oboru LIT. Pro porovnání s úspěšností ve vyhledávání jednoslovných termínů viz obrázek 3.1.

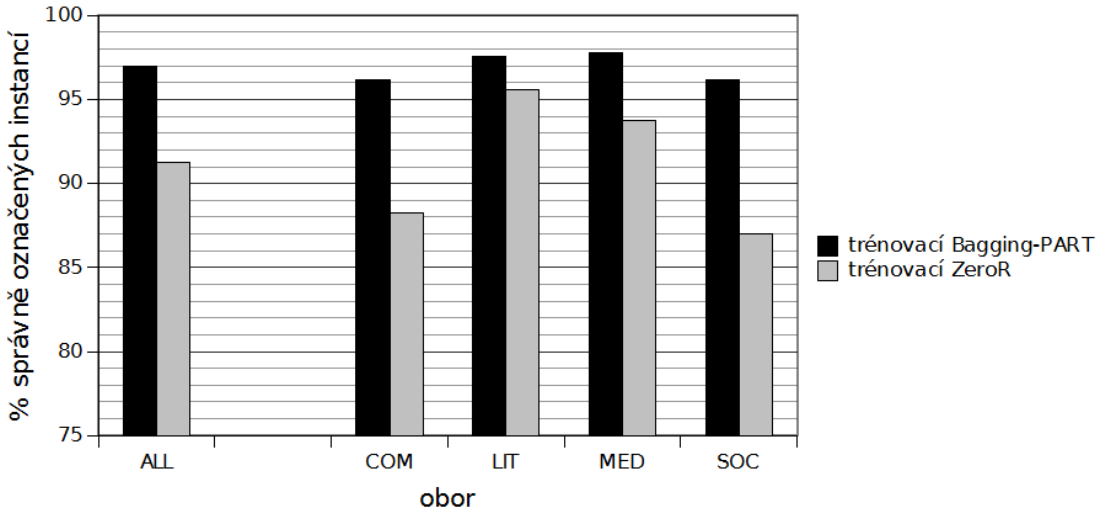
3.2.2 Automatické vyhledávání v testovacích datech

Parametry experimentu 4: Srovnání úspěšnosti automatického vyhledání víceslovných termínů v testovacích datech. V experimentu 4 je srovnávána úspěšnost označování víceslovných termínů pomocí metod Bagging-PART a ZeroR v jednotlivých oborech v testovacích datech. Materiál je ve formě bigramů, bez lemmatizace, s opakováním a bez rozlišení slovních druhů. V experimentu byly použity všechny vlastnosti určené pro vyhledávání víceslovných termínů (označené jako MWT). Hranice mezi termíny a netermíny je v nástroji Weka defaultně nastavena na hodnotu 0,5.

Z důvodu výpočetní náročnosti[4] jsou testovací data pro identifikaci víceslovných termínů daleko méně rozsáhlá než u termínů jednoslovných: víceslovné termíny byly automaticky vyhledány pouze ve čtyřech disciplínách (COM, LIT, MED, SOC) v celkem 10 tisících textových pozic.

[4]Výpočty trvají i na poměrně malých datech i několik dnů.

Obrázek 3.6: Úspěšnost (hodnota míry *accuracy*) v trénovacích datech ze čtyř oborů (COM, LIT, MED, SOC), víceslovné termíny, metoda Bagging-PART. Výsledky je třeba srovnat s úspěšností metody ZeroR. Úspěšnost v trénovacích datech bývá obvykle vyšší, protože tatáž data jsou použita k natrénování i k následnému automatickému vyhledání.



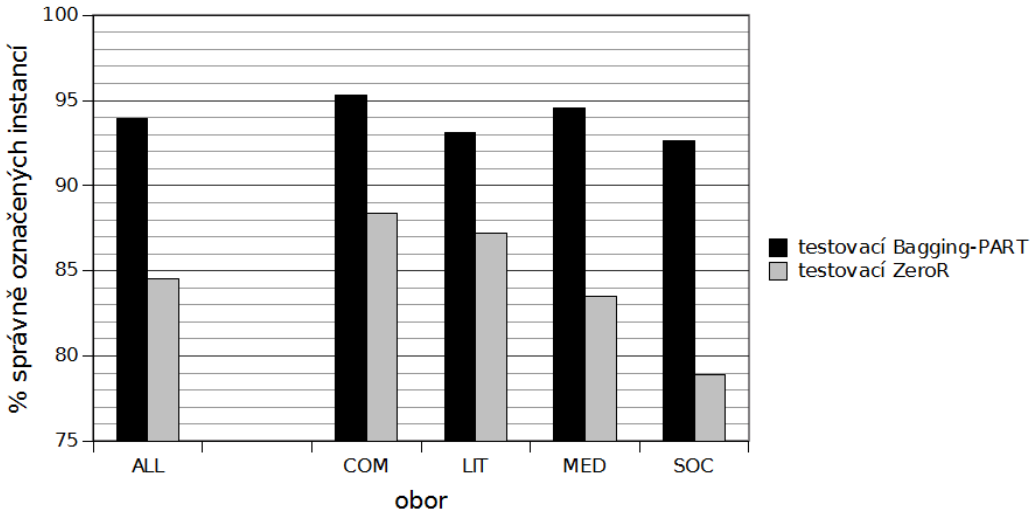
Úspěšnost (*accuracy*) vyhledávání víceslovných termínů v testovacích datech je nižší než u dat trénovacích: průměrně asi 94 % bigramů je správně zařazeno mezi víceslovné termíny nebo netermíny (oproti průměrným 97 % u trénovacích dat, viz kap. 3.2.1). Vyšší úspěšnost v trénovacích datech je obvyklá a je dána tím, že metoda je natrénována na konkrétních bigramech obsažených v trénovacích datech.

V obrázku 3.7 jsou obsaženy výsledky metody Bagging-PART a srovnávací metody ZeroR. Nejvyšší úspěšnost sice metoda vykazuje u oboru COM (cca 95 %), největší rozdíl oproti ZeroR je však u oboru SOC (téměř 12% zlepšení oproti 7 % u oboru COM).

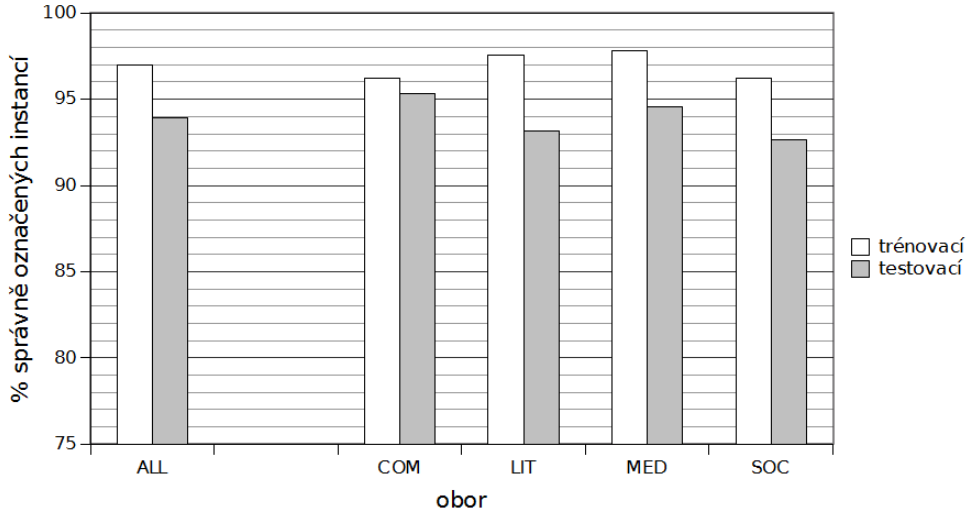
V obrázku 3.8 je znázorněn rozdíl úspěšnosti vyhledávání termínů v trénovacích a testovacích datech.

Při hodnocení úspěšnosti je možné se zaměřit jen na příznakové členy, v tomto případě na víceslovné termíny (pomocí evaluačních měr *precision* a *recall* a z nich odvozené míry *F-measure*). Z obrázku 3.9 je zřejmé, že v rámci testovacích dat vykazuje nejlepší výsledky ze čtyř zkoumaných oborů obor MED s nejvyšší hodnotou *precision* a *F-measure*. Nejvyšší *recall* je v oboru COM, ovšem za cenu poměrně nízké hodnoty *precision*. Nejméně úspěšné je vyhledávání termínů v oboru LIT.

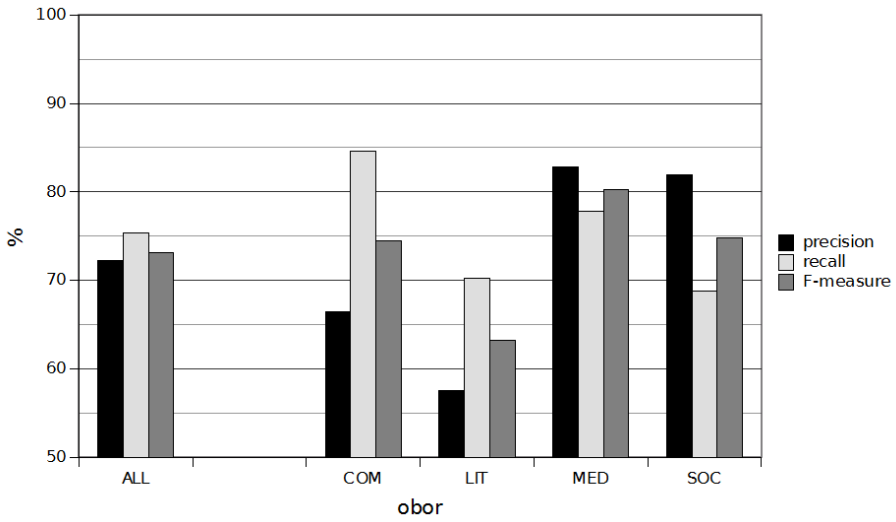
Obrázek 3.7: Úspěšnost (hodnota míry *accuracy*) v testovacích datech ze čtyř oborů (COM, LIT, MED, SOC), víceslovné termíny, metoda Bagging-PART. Výsledky je třeba srovnat s úspěšností metody ZeroR (viz kap. 2.4).



Obrázek 3.8: Úspěšnost vyhledávání víceslovných termínů (hodnota míry *accuracy*) v trénovacích a testovacích datech ze čtyř oborů (COM, LIT, MED, SOC), metoda Bagging-PART. Ve všech případech je úspěšnost vyšší u trénovacích dat, což je obvyklé, protože metoda je natrénována na konkrétních bigramech obsažených v trénovacích datech.



Obrázek 3.9: Hodnoty evaluačních *měr precision, recall* a *F-measure* v testovacích datech ze čtyř oborů (COM, LIT, MED, SOC), víceslovné termíny, metoda Bagging-PART. *Precision* vyjadřuje, kolik procent z označených bigramů jsou skutečně víceslovné termíny, hodnota *recall* odpovídá procentu víceslovných termínů, které byly skutečně vyhledány. *F-measure* je průměr měr *precision* a *recall*.

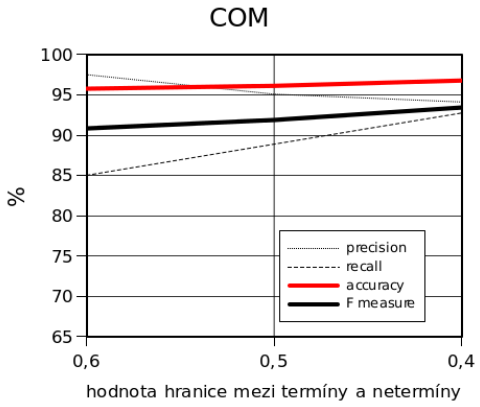


3.3 Hodnota terminologické platnosti

V terminologii, stejně jako v dalších oblastech lingvistiky, lze využít princip škály - některé termíny jsou silnější než jiné, některá slova lze s větší jistotou zařadit mezi netermíny (více o škále v kap. 1.3). V předkládaném výzkumu se pracuje s terminologickou platností jednotlivých textových pozic (resp. bigramů) na škále 0 až 1 (hodnotu 0 mají nejsilnější netermíny, hodnotu 1 nejsilnější termíny). Hranice mezi termíny a netermíny je nastavena na hodnotu 0,5 (ta je v data-miningovém nástroji Weka v defaultním nastavení a všechny experimenty jsou optimalizovány pro tuto hodnotu; Hall, 2009).

Hodnota 0,5 je sice intuitivní hranicí mezi termíny a netermíny, ale je třeba na ni pohlížet jako na danou či neměnnou - podle různých výzkumných úkolů lze hranici posunovat, a výběr automaticky vyhledaných termínů tím zpřísnit, nebo naopak pravidla zmírnit. Následující experimenty zkoumají výhodnost různých hodnot, na něž lze hranici mezi termíny a netermíny umístit (0,4, 0,5 a 0,6). Vyhodnocení se zaměřuje na evaluační míry zkoumající úspěšnost a přesnost vyhledávání termínů (popř. i netermínů) v datech: *accuracy, precision, recall* a *F-measure*. Největší vliv mají změny zkoumaných hodnot na míry precision a recall - čím výše je hranice nastavená, tím nižší je *recall* (počet nalezených termínů ze všech termínů obsažených v datech), ale tím vyšší je *precision* (počet automaticky označených termínů, které skutečně jsou termíny). Záleží tedy na tom, zda chceme

Obrázek 3.10: Hodnota evaluačních měr *accuracy, precision, recall* a *F-measure* při různých nastaveních hranice mezi jednoslovnými termíny a netermíny. Trénovací data: obor informatika (COM); použitá metoda Bagging-PART.



najít co největší počet termínů, nebo zda chceme, aby výsledky byly co nejpřesnější bez ohledu na nižší procento nalezených termínů.

Tato práce se však nesoustředí jen na úspěšnost metody nebo na automaticky vyhledané termíny, ale i na vlastnosti, které při vyhledávání termínů hrají největší roli. Proto je ve všech experimentech zachována původní hodnota hranice 0,5.

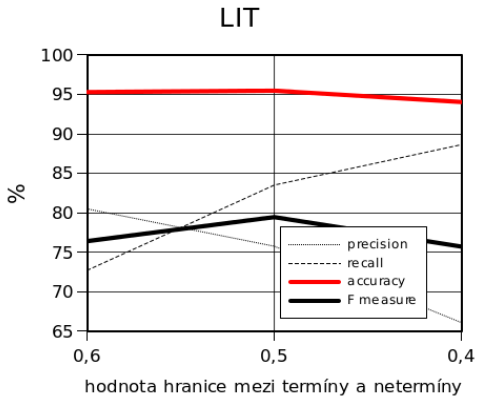
3.3.1 Hranice mezi jednoslovnými termíny a netermíny

Parametry experimentu 5: Hranice mezi jednoslovnými mezi termíny a netermíny. Experiment je zaměřen na určení hranice mezi jednoslovnými termíny a netermíny na škále terminologické platnosti 0 až 1. Zkoumají se tři možné hodnoty: 0,4, 0,5 a 0,6. Pro každý z oborů v trénovacím materiálu (COM, LIT, MED a SOC a ve sloučených datech ALL) jsou vypočítány evaluační míry *accuracy, precision, recall* a *F-measure* (kap. 2.4). Materiál je v podobě nelemmatizované, s opakováním a bez slovních druhů. V experimentu byly použity pouze čtyři z vlastností: RFQdiscRFQcompar, ARF, RDist, a SDRD (viz kap. 5.1.3).

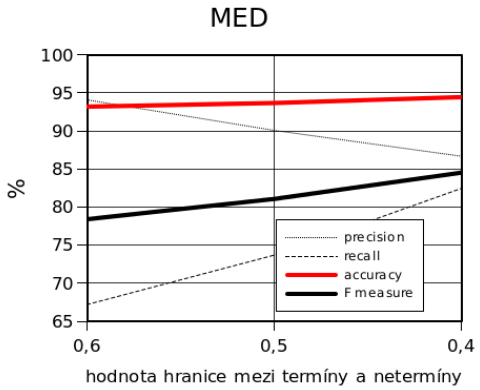
V experimentech je všem textovým pozicím automaticky přidělována hodnota terminologické platnosti na škále 0 až 1. Textové pozice s hodnotou přesahující stanovenou mez jsou považovány za termíny, ostatní textové pozice jsou zařazeny mezi netermíny. Výsledky se hodnotí srovnáním s ručně označenými daty (které z ručně označených termínů byly zařazeny mezi netermíny, a naopak).

V obrázcích 3.10 až 3.13 jsou pro čtyři obory (COM, LIT, MED a SOC) znázorněny hodnoty evaluačních měr. Díky nim lze zkoumat výhodnost jednotlivých hodnot hranice mezi termíny a netermíny (0,4, 0,5, 0,6). Úspěšnost (*accuracy*) zařazování textových pozic mezi termíny či netermíny zůstává poměrně stabilní u všech 4 oborů, u oborů COM a MED je možné vypozorovat určitou tendenci ke zlepšení při snižování hodnoty hranice.

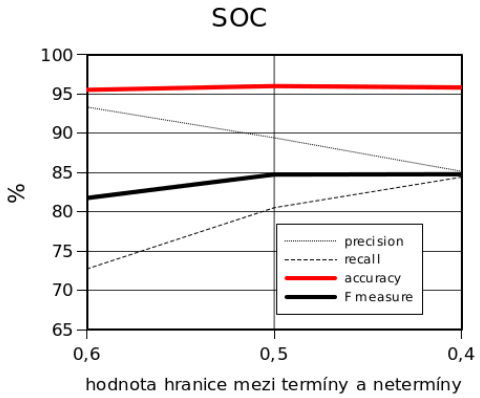
Obrázek 3.11: Hodnota evaluačních měr *accuracy, precision, recall* a *F-measure* při různých nastaveních hranice mezi jednoslovnými termíny a netermíny. Trénovací data: obor literatura (LIT); použitá metoda Bagging-PART.



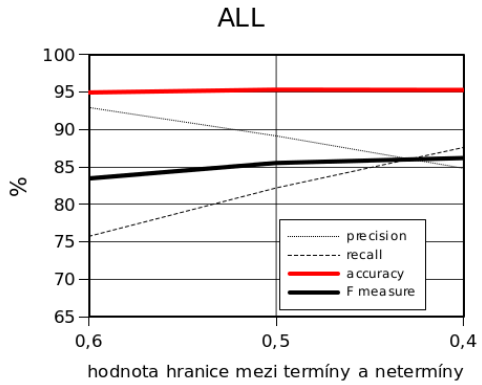
Obrázek 3.12: Hodnota evaluačních měr *accuracy, precision, recall* a *F-measure* při různých nastaveních hranice mezi jednoslovnými termíny a netermíny. Trénovací data: obor medicína (MED); použitá metoda Bagging-PART.



Obrázek 3.13: Hodnota evaluačních měr *accuracy, precision, recall* a *F-measure* při různých nastaveních hranice mezi jednoslovnými termíny a netermíny. Trénovací data: obor sociologie (SOC); použitá metoda Bagging-PART.



Obrázek 3.14: Hodnota evaluačních měr *accuracy, precision, recall* a *F-measure* při různých nastaveních hranice mezi jednoslovnými termíny a netermíny. Trénovací data: sloučená data ze 4 oborů (ALL); použitá metoda Bagging-PART.



Poměrně velké změny (až o desítky procent) jsou viditelné u měr *precision* a *recall*, přičemž *precision* klesá se snižující se hranicí mezi termíny a netermíny, kdežto *recall* naopak stoupá.

Hodnota *F-measure* (průměr hodnot *precision* a *recall*) přírodovědných/technických oborů stoupá se snižující se hodnotou hranice, kdežto u humanitních oborů se defaultní hranice 0,5 jeví jako optimální.

Obrázek 3.14 shrnuje údaje z dat všech čtyř testovacích oborů. Na rozdíl od jednotlivých oborů je hodnota *accuracy* i *F-measure* poměrně vyvážená u všech tří nastavení. Největší rozdíl je opět mezi mírami *precision* a *recall*.

Pokud by hranice mezi termíny a netermíny byla předmětem zájmu (např. při výzkumu se specifickým cílem), bylo by možné hranici mezi termínem a netermínem nastavit podle hodnot evaluačních měr i podle typu zkoumaných oborů (pro humanitní obory je vhodnější hranice 0,5, pro přírodovědné a technické by bylo výhodné posunout hranici na hodnotu 0,4). Protože předkládaná práce se nesoustředí specificky na automatické vyhledání termínů (ale spíše na to, co se o termínech můžeme na základě metody TERMIT dozvědět), zůstane při všech experimentech hranice mezi termíny a netermíny nastavena na původní hodnotě 0,5.

V trénovacích datech, kde byly ručně označeny všechny textové pozice, se při nastavení hranice na hodnotu 0,5 nevyhledají termíny jako: *počítačům* (COM), *Čtenář* (LIT), *kardiochirurgie* (MED) a *mravní* (SOC) (termíny, kterým byla automaticky přidělena hodnota terminologické platnosti nižší než 0,5). Naopak za termíny jsou automaticky označeny následující netermíny: *požadovaný* (COM), *stávající* (LIT), *zhoršují* (MED) a *ztotožňované* (SOC) (netermíny, jimž byla automaticky přidělena hodnota terminologické platnosti vyšší než 0,5). Úplný seznam textových pozic s hodnotami terminologické platnosti je v příloze D.

3.3.2 Hranice mezi víceslovnými termíny a netermíny

Parametry experimentu 6: Hranice mezi víceslovnými termíny a netermíny. Experiment je zaměřen na určení hranice mezi termíny a netermíny na škále terminologické platnosti 0 až 1. Zkoumají se tři možné hodnoty: 0,4, 0,5 a 0,6. Pro každý z oborů v trénovacím materiálu (COM, LIT, MED a SOC) jsou vypočítány evaluační míry *accuracy, precision, recall* a *F-measure* (kap. 2.4). Materiál je v podobě bigramů, bez lemmatizace, s opakováním a bez slovních druhů. V experimentu bylo použito šest vybraných vlastností: MWT:T1, MWT:T2, MWT:t-score, MWT:MI-score, MWT:Oblig a MWT:Prox (viz kap. 5.2.3).

Na základě metody Bagging-PART byla všem bigramům v trénovacích datech automaticky přidělena hodnota na škále 0 až 1, kde 0 je nejslabší a 1 nejsilnější terminologická platnost. Bigramy s hodnotou přesahující stanovenou mez jsou považovány za víceslovné termíny, ostatní textové pozice jsou zařazeny mezi netermíny. Výsledky se porovnávají s ručně označenými daty.

Stejně jako u jednoslovných termínů, i zde je velmi důležité, jaké jsou cíle výzkumu. Pokud je úkolem najít co největší počet víceslovných termínů (i za cenu nižší kvality některých z nich), pak je třeba nastavit hranici mezi termíny a netermíny na nižší hodnotu. Pokud naopak budeme chtít co největší jistotu, že označené bigramy jsou skutečně víceslovné termíny, a nezáleží na tom, že některé termíny zůstanou nenalezeny, pak hranici posuneme výše.

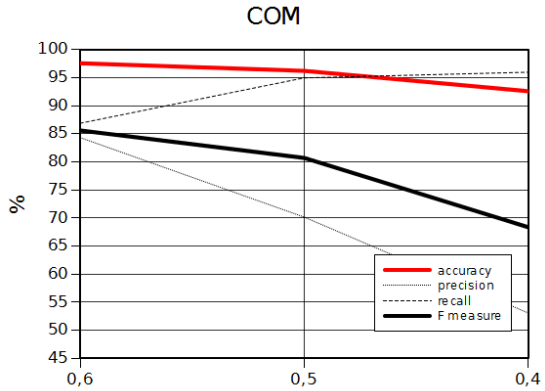
Celkově se ale zdá u většiny oborů (snad s výjimkou MED), že u víceslovných termínů by posunutí hranice mezi termíny a netermíny na hodnotu 0,6 zlepšilo celkovou úspěšnost (*accuracy*) zařazení bigramů mezi víceslovné termíny či netermíny i kvalitu vyhledaných víceslovných termínů (*precision*), ovšem za cenu zvýšení počtu nenalezených termínů.

Naopak posunutím hranice na hodnotu 0,4 by se zásadně zlepšil *recall* - u všech oborů by byl velmi vysoký, zhruba 95 % - naprostá většina víceslovných termínů se tedy nachází nad touto hranicí - v tomto případě by ale mezi termíny byl zařazen i velký počet bigramů, které ve skutečnosti termíny nejsou.

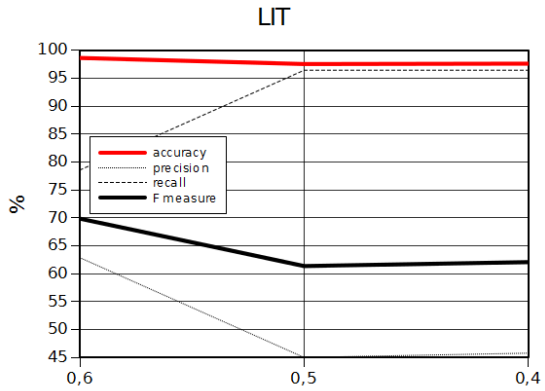
Stejně jako u jednoslovných termínů, ani v tomto případě však změna hranice neznamená tak zásadní zlepšení, aby bylo třeba manipulovat s defaultním nastavením data-miningového nástroje Weka (cíle této práce to nevyžadují). Proto i zde zůstane hranice mezi termíny a netermíny na původní hodnotě 0,5.

V datech ze čtyř zkoumaných oborů (COM, LIT, MED, SOC) se při nastavení hranice na hodnotu 0,5 nevyhledají termíny jako: *optimalizace obvodů* (COM), *narativní kompetenci* (LIT), *Srdeční frekvence* (MED) a *občanského statusu* (SOC) (termíny, kterým byla automaticky přidělena hodnota terminologické platnosti nižší než 0,5). Naopak za termíny jsou automaticky označeny následující netermíny: *neprodlužovala instrukce* (COM), *vyprávění*

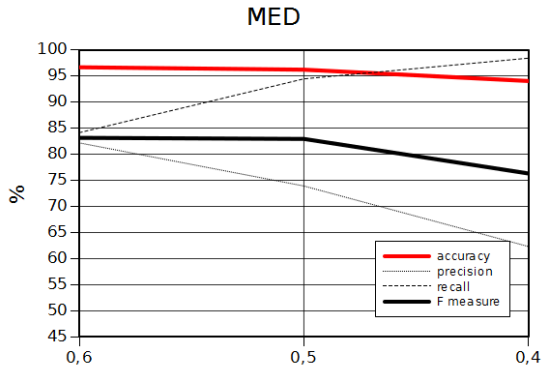
Obrázek 3.15: Hodnota evaluačních měr *accuracy, precision, recall* a *F-measure* při různých nastaveních hranice mezi víceslovnými termíny a netermíny. Trénovací data: obor informatika (COM); použitá metoda Bagging-PART.



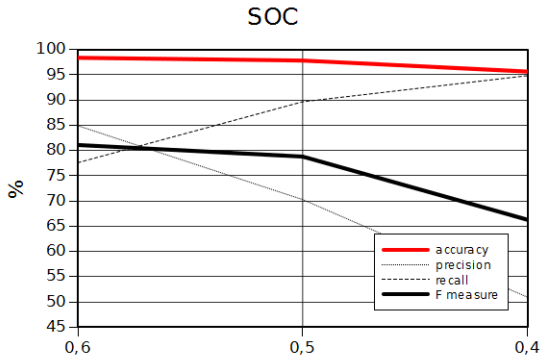
Obrázek 3.16: Hodnota evaluačních měr *accuracy, precision, recall* a *F-measure* při různých nastaveních hranice mezi víceslovnými termíny a netermíny. Trénovací data: obor literární věda (LIT); použitá metoda Bagging-PART.



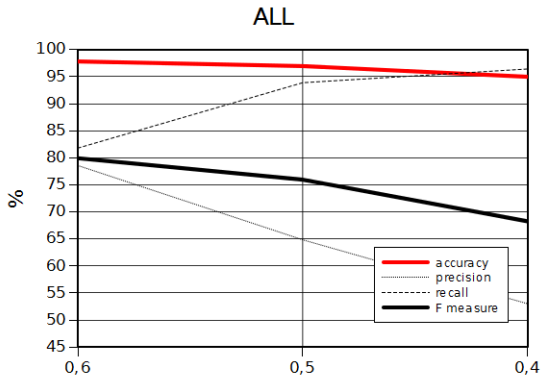
Obrázek 3.17: Hodnota evaluačních měr *accuracy, precision, recall* a *F-measure* při různých nastaveních hranice mezi víceslovnými termíny a netermíny. Trénovací data: obor lékařství (MED); použitá metoda Bagging-PART.



Obrázek 3.18: Hodnota evaluačních měr *accuracy, precision, recall* a *F-measure* při různých nastaveních hranice mezi víceslovnými termíny a netermíny. Trénovací data: obor sociologie (SOC); použitá metoda Bagging-PART.



Obrázek 3.19: Hodnota evaluačních měr *accuracy, precision, recall* a *F-measure* při různých nastaveních hranice mezi víceslovnými termíny a netermíny. Trénovací data: všechny obory; použitá metoda Bagging-PART.



vnímáme (LIT), *popsáno nadýmání* (MED) a *potřebou nalézt* (SOC) (netermíny, jimž byla automaticky přidělena hodnota terminologické platnosti vyšší než 0,5). Úplný seznam textových pozic s hodnotami terminologické platnosti je v příloze D.

3.4 Shrnutí

V rámci trénovacích dat pro jednoslovné termíny je úspěšnost (hodnota míry *accuracy*) více než 95 % (tj. 95 % textu zařazeno správně mezi termíny nebo netermíny). V testovacích datech ze všech 37 oborů v korpusu SYN 2010 dosáhla metoda TERMIT vysokých hodnot evaluačních měr *precision* a *recall*, 85 % a 81 % (to znamená, že 85 % automaticky označených termínů jsou skutečně termíny a 81 % termínů vyskytujících se v daném textu bylo skutečně nalezeno.

V rámci trénovacích dat pro víceslovné termíny je úspěšnost (hodnota míry *accuracy*) průměrně 97 % (tj. 97 % procent textu zařazeno správně mezi termíny nebo netermíny). V testovacích datech ze čtyř oborů metoda TERMIT dosáhla uspokojivých hodnot evaluačních měr *precision* a *recall* 72 % a 75 % (tzn. 72 % automaticky označených termínů jsou skutečně termíny a 75 % termínů bylo skutečně nalezeno).

Vysoká úspěšnost automatického vyhledávání jednoslovných termínů je podstatná ze dvou důvodů: 1. na základě úspěšnější metody lze vyvozovat s větší jistotou závěry o termínu jako základní jednotce terminologie, např. jaké jsou vlastnosti termínů v reálných textech nebo jaký je počet jednoslovných termínů v různých oborech, a 2. výsledků úspěšné metody automatického vyhledávání termínů je možné využít i prakticky, k vytváření terminologických slovníků, k automatické indexaci, ke strojovému překladu atd. (kap. 1.1.2).

Hranice mezi termíny a netermíny je v rámci experimentů nastavena na hodnotu 0,5, ač u některých oborů či s ohledem na některé výzkumné cíle by mohlo být výhodnější ji nastavit na hodnotu vyšší, či nižší. Takový posun by měl vliv především na precision a recall vyhledávání termínů, kdežto na celkovou úspěšnost (accuracy) by změna měla spíše menší vliv.

4 Automaticky vyhledané termíny

Kapitola se věnuje analýze termínů automaticky vyhledaných metodou TERMIT. V první řadě se zde porovnává počet termínů jednoslovných a víceslovných termínů v textech akademických disciplín, a dále počet jednoslovných termínů v textech i slovní zásobě velkého množství oborů, které jsou k dispozici v korpusu SYN2010.

Na základě sdílení automaticky vyhledaných jednoslovných termínů dvěma nebo více disciplínami lze mezi nimi vyhledávat souvislosti či určovat příbuznost. Druhá část kapitoly se zabývá právě tím, jak silné jsou vazby mezi jednotlivými obory.

V závěru kapitoly je pozornost soustředěna na termíny, kterým byla automaticky přidělena nejvyšší terminologická platnost, a současně na netermíny, které mají hodnotu terminologické platnosti nejnižší. Zároveň jsou zmíněny některé charakteristické vlastnosti termínů a terminologií čtyř blíže zkoumaných oborů, tedy informatiky, literární vědy, lékařství a sociologie.

4.1 Počet automaticky vyhledaných termínů

4.1.1 Srovnání počtu jednoslovných a víceslovných termínů

Jednou z častých otázek týkajících se termínů je počet jednoslovných a víceslovných termínů v textech a ve slovní zásobě jednotlivých oborů. Zde se budeme věnovat především počtu termínů ve čtyřech blíže zkoumaných oborech (COM, LIT, MED a SOC)[1].

U jednoslovných termínů vždy záleží na tom, zda bereme v úvahu pouze ty samostatně stojící, nebo všechny jednoslovné termíny včetně těch, které se podílejí na víceslovných termínech. V tabulkách i obrázcích jsou tyto dva případy vždy pečlivě označeny.

U víceslovných termínů je třeba vyřešit otázku, zda sledovat počet samotných termínů, nebo počet textových pozic, které obsazují. Protože existují víceslovné termíny složené z více než dvou textových pozic[2], nelze jejich počet zjistit automaticky, ale vždy jen ručním zpracováním.

[1] Počet jednoslovných termínů ve 37 různých oborech je zachycen v kap. 4.1.2, a to především v tabulce 4.5.

[2] Šestislovný termín může být započítán pouze jednou, ale přitom obsazuje šest textových pozic, na druhou stranu může dojít k situaci, kdy dva dvouslovné termíny následují v textu těsně za sebou, obsazují tedy 4 textové pozice - tyto dva případy nelze odlišit automaticky.)

Tabulka 4.1: Počet automaticky vyhledaných jednoslovných a víceslovných termínů v trénovacích a testovacích datech ze 4 oborů (COM, LIT, MED, SOC) a v testovacích datech ze všech oborů dostupných v korpusu SYN2010 (37 oborů). Počet samostatně stojících jednoslovných termínů lze pouze odhadovat, na základě analýzy ručně označených dat tvoří zhruba 60 % jednoslovných termínů (viz tab. 4.2). Víceslovné termíny nebylo možno z důvodu výpočetní náročnosti vyhledávat v rozsáhlých datech všech akademických oborů. Všechny údaje jsou převzaty z experimentů 7, 9 a 10 v kap. 4.1.2, 4.3 a 4.4.

	SWT samost. (odhad)	MWT	SWT všechny
trénovací (4 obory)	9.5	13.1	15.8
testovací (4 obory)	10.6	13.8	17.6
testovací (37 oborů)	13	x	21.7

4.1.1.1 Počet jednoslovných a víceslovných termínů v textech

Procento jednoslovných termínů v textech akademických oborů se pohybuje zhruba kolem 16 až 22 % ze všech textových pozic (tab. 4.1, poslední sloupec), přibližně 60 % z nich je samostatně stojících, zbylých 40 % jednoslovných termínů je součástí termínů víceslovných (viz tab. 4.2). Ve velkém počtu akademických textů z celkem 37 oborů je průměrné procento textových pozic obsazených jednoslovnými termíny (samostatnými i nesamostatnými) 22 %, přičemž tato hodnota kolísá mezi 9 a 47 % (viz tab. 4.5 v kap. 4.1.2).

Počet textových pozic obsazovaných víceslovnými termíny se pohybuje mezi 13 a 14 % ze všech textových pozic (tab. 4.1, prostřední sloupec). Počet víceslovných termínů (tedy nikoli textových pozic, ale skutečně termínových kolokací) je v řádech desítek termínů (55 až 106 termínů) v textu o tisíci pozicích (viz tab. 4.3). Velké výchylky jsou způsobeny odlišností zkoumaných akademických disciplín. Nejméně víceslovných termínů je oboru literární věda, kde je ostatně nejmenší počet veškerých termínů ze čtyř blíže zkoumaných oborů, nejvyšší počet víceslovných termínů je v oboru lékařství.

Počet samostatně stojících jednoslovných termínů v textech je ve všech zkoumaných případech vyšší než počet víceslovných termínů (jde skutečně o počet, nikoli o textové pozice), a to mnohdy i dvojnásobně (viz tab. 3). Údaje o poměru terminologických kolokací ve slovní zásobě nejsou k dispozici, nicméně lze předpokládat, že počet víceslovných termínů bude mnohonásobně převyšovat počet termínů jednoslovných (mj. díky kombinatorickým možnostem víceslovných termínů složených z jednoslovných, srov. *aortální/ mitrální/ umělá/ srdeční/ trikuspidální chlopeň* a *aortální krev/ váček/ stenóza/ oblouk*).

Na základě ručně označených trénovacích dat lze odhadnout, kolik procent víceslovných

Tabulka 4.2: Procento samostatně stojících jednoslovných termínů v ručně označených trénovacích datech. Za samostatně stojící termíny jsou považovány ty, které v textu nejsou součástí termínu víceslovného. Údaje jsou převzaty z experimentu 9 v kap. 4.3.

	SWT všechny pozice	SWT samostatné	v %
COM	513	329	64.1
LIT	241	181	75.1
MED	503	236	46.9
SOC	304	161	53
ALL	1561	907	**58.1**

Tabulka 4.3: Srovnání průměrného počtu automaticky vyhledaných jednoslovných a víceslovných termínů textu o tisíci textových pozic (vč. interpunkce). Z jednoslovných jsou zde brány v úvahu pouze ty samostatně stojící, u víceslovných je zaznamenán počet celých termínů, tedy nikoli textových pozic. Všechny údaje jsou převzaty z experimentů 7, 9 a 10 v kap. 4.1.2, 4.3 a 4.4.

	SWT trénovací	**MWT** trénovací	SWT testovací	**MWT** testovací
COM	128	**75**	132	**48**
LIT	55	**39**	73	**39**
MED	122	**86**	105	**71**
SOC	72	**44**	66	**85**
ALL průměr	94	**61**	106	**68**

termínů je tvořeno termínovými textovými pozicemi. Drtivá většina víceslovných obsahuje alespoň jeden termín jednoslovný, v některých případech dva i více (k tomu viz i kap. 4.4). S jistotou tedy můžeme tvrdit, že minimálně 50 % textových pozic obsazených víceslovnými termíny je tvořeno jednoslovnými. Ve skutečnosti je ale toto procento daleko vyšší, jde zhruba o tři čtvrtiny, v některých případech dokonce téměř 85 % (viz tab. 4.4).

V tabulce 4.3 je možné porovnat jednotlivé obory co do počtu jednoslovných a víceslovných termínů. Humanitní obory literární věda a sociologie obecně mají výrazně nižší počet termínů v textech, a to jak jednoslovných, tak víceslovných[3]. Celkově nejnižší počet termínů je v literární vědě, kdežto nejvíce termínů se vyskytuje v oborech informatika a lékařství (u obou je počet termínů vyrovnaný).

[3]K podobným závěrům lze dospět i na základě experimentu 7 v kap. 4.1.2.

Tabulka 4.4: Počet textových pozic víceslovných termínů, které jsou tvořeny termíny jednoslovnými. Naprostá většina víceslovných termínů je tvořena alespoň jedním, ale mnohdy i několika jednoslovnými. Údaje jsou převzaty z experimentu 10 v kap. 4.4.

	MWT txt pozice	z toho SWT	v %
COM	236	184	78.0
LIT	89	60	67.4
MED	315	267	84.8
SOC	203	143	70.4
ALL	843	654	**77.6**

4.1.1.2 Počet ručně a automaticky vyhledaných termínů

Při posuzování počtu termínů, které je založeno na automaticky vyhledávaných termínech, se nabízí otázka, do jaké míry je údaj o množství jednoslovných a víceslovných termínů spolehlivý. V kapitole o úspěšnosti automatické identifikace bylo na základě statistických evaluačních měr stanoveno, jaké procento textových pozic je automaticky správně zařazeno mezi termíny a netermíny (viz kap. 3), nicméně o přesnosti počtu vyhledaných termínů to vypovídá jen málo. Proto je třeba provést analýzu toho, do jaké míry si odpovídají údaje o množství termínů v ručně a automaticky zpracovaných datech.

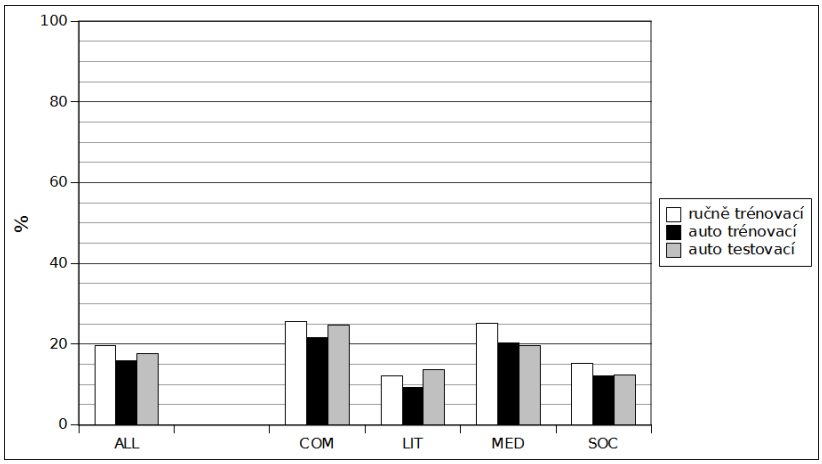
V obrázcích 4.1 a 4.2 je srovnán počet ručně a automaticky vyhledaných jednoslovných a víceslovných termínů. Z obou obrázků je zřejmé, že množství termínů je poměrně vyrovnané v obou typech dat (ručně i automaticky zpracovaných). Zvláště to platí o termínech jednoslovných, kde se rozdíly pohybují maximálně okolo tří procent. Důvodem této přesnosti počtu termínů je to, že některé nevyhledané termíny (FN, viz kap. 2.4) jsou nahrazeny nesprávně označenými netermíny (FP, tamtéž) - metoda TERMIT tak vyrovnává odchylky stejnoměrně oběma směry.

Z obou obrázků je možno vyvodit závěr, že údaje o počtu termínů v textech, zvláště pak termínů jednoslovných, jsou velmi přesné a spolehlivé.

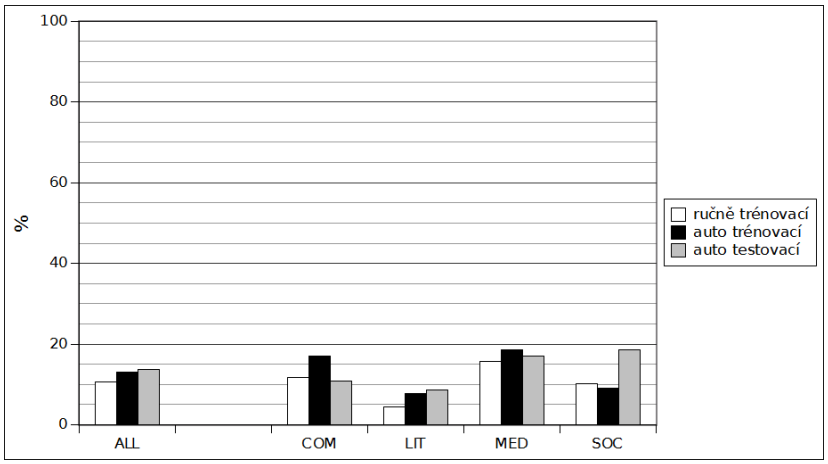
4.1.2 Jednoslovné termíny v různých oborech

Parametry experimentu 7: Počet automaticky vyhledaných termínů v testovacích datech. V experimentu byly za pomoci metody Bagging-PART automaticky označeny jednoslovné termíny v testovacích datech (37 oborů z korpusu SYN2010 s různým počtem textových pozic v každém oboru). Na základě vyhledaných termínů byly porovnány počty termínů v jednotlivých oborech. Materiál je v podobě nelemmatizované, s opakováním a bez slovních druhů. V experimentu byly použity čtyři vlastnosti: RFQdiscRFQcompar, ARF, RDist, a SDRD (viz kap. 5.1.3). Hranice mezi termíny a netermíny je

Obrázek 4.1: Počet ručně a automaticky vyhledaných jednoslovných termínů v trénovacích a testovacích datech.



Obrázek 4.2: Počet ručně a automaticky vyhledaných víceslovných termínů v trénovacích a testovacích datech.



v nástroji Weka defaultně nastavena na hodnotu 0,5 (viz kap. 3.3.1).

Automatické vyhledání termínů v textech velkého počtu akademických disciplín umožňuje srovnání oborů: Na jeho základě lze zjistit, kolik procent textu v různých oborech obsazují termíny, kolik procent ze všech užitých lemmat tvoří termíny, nebo zda se tato procenta liší v různých typech oborů.

V textech humanitních (vč. sociálněvědných) oborů je obecně méně jednoslovných termínů než v textech oborů přírodovědných a technických[4]; je to zřejmé jak z počtu termínů v trénovacích datech ze čtyř vybraných oborů (viz tab. 4.1 a 4.3), tak z počtu automaticky vyhledaných termínů v textech všech oborů dostupných v korpusu SYN2010 (viz tab. 4.5).

Pomocí metody Bagging-PART byly automaticky vyhledány jednoslovné termíny v 37 akademických oborech z korpusu SYN2010[5] (viz tab. 2.1). Pro každou textovou pozici byla vypočtena hodnota terminologické platnosti na škále mezi hodnotami 0 a 1 - mezi termíny byly zařazeny ty textové pozice, které dosáhly hodnoty terminologické platnosti 0,5 a vyšší[6].

V tabulce 4.5 a v obrázcích 4.3 a 4.4 se porovnává velké množství akademických disciplín na základě dvou údajů:

1. množství termínů v textech jednotlivých oborů, tj. procento textových pozic obsazených automaticky vyhledanými termíny
2. rozsah terminologie ve slovní zásobě oborů, tj. procento automaticky označených terminologických lemmat ze všech použitých lemmat.

Z tabulky vyplývá, že zhruba 22 % akademických textů je tvořeno termíny a že průměrně 31 % lemmat v akademických oborech jsou termíny. V humanitních disciplínách je přitom oproti oborům přírodovědným/technickým daleko nižší procentuální zastoupení termínů jak v textech, tak mezi lemmaty. Například v oboru psychologie je jen zhruba 10 % termínů v textech a 16 % lemmat je terminologických, kdežto v botanice je více než třetina textových pozic termínových a až polovinu lemmat tvoří termíny.

[4]Rozdělení oborů na humanitní, sociálněvědné, přírodovědné, technické, případně ještě formální, aplikované apod. není všeobecně sjednocené, proto i zde z praktických důvodů rozlišujeme pouze dvě skupiny oborů: 1. humanitní (zahrnující i sociálněvědné) a 2. přírodovědné a technické.

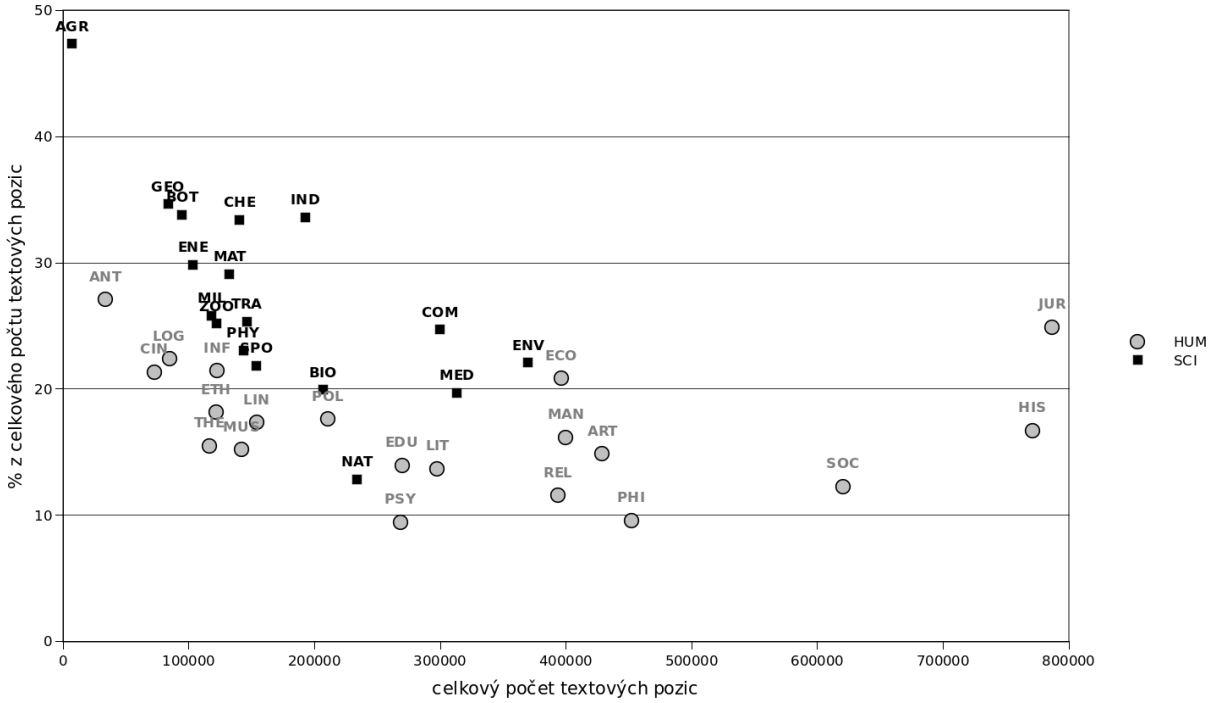
[5]Zhodnocení úspěšnosti založené na ruční analýze náhodně vybraných úseků z každého oboru ukazuje, že automatické vyhledávání termínů napříč množstvím různých oborů je poměrně úspěšné, průměrná hodnota míry *accuracy* pro všechny obory je 91,1 % (více v kap. 3.1.2).

[6]Hranici mezi termíny a netermíny se věnuje kap. 3.3.1.

Tabulka 4.5: Procento termínů v textech (procento termínových textových pozic mezi všemi textovými pozicemi) a ve slovní zásobě (procento terminologických lemmat z celkového počtu jedinečných lemmat) ve 37 oborech dostupných v SYN2010. Údaje jsou založeny na automatickém označení jednoslovných termínů.

Obor	**% T v textech**	**Obor**	**% T lemmat**
průměr	**21.66**	**průměr**	**30.97**
Humanitní		**Humanitní**	
PSY	9.44	PSY	16.23
PHI	9.57	SOC	18.28
REL	11.53	POL	19.07
SOC	12.25	EDU	20.83
LIT	13.66	HIS	21.18
EDU	13.93	THE	21.66
ART	14.86	MAN	21.75
MUS	15.16	PHI	21.90
THE	15.48	MUS	22.28
MAN	16.16	JUR	22.82
HIS	16.69	REL	23.80
LIN	17.33	ETH	25.21
POL	17.59	ECO	25.30
ETH	18.14	ART	25.90
ECO	20.82	LIT	26.58
CIN	21.32	LOG	27.56
INF	21.42	LIN	28.00
LOG	22.35	CIN	29.85
JUR	24.89	ANT	34.24
ANT	27.08	INF	36.56
průměr	**16.98**	**průměr**	**24.45**
Přír. a tech.		**Přír. a tech.**	
NAT	12.82	MIL	21.48
MED	19.66	NAT	27.69
BIO	19.90	SPO	31.31
SPO	21.81	PHY	32.48
ENV	22.05	ZOO	33.42
PHY	23.03	TRA	33.71
COM	24.67	MED	34.50
ZOO	25.14	BIO	35.76
TRA	25.32	ENV	36.56
MIL	25.80	MAT	36.90
MAT	29.07	ENE	38.03
ENE	29.83	COM	44.58
CHE	33.36	IND	46.15
IND	33.58	CHE	50.83
BOT	33.80	AGR	50.92
GEO	34.61	BOT	51.04
AGR	47.31	GEO	51.72
průměr	**27.16**	**průměr**	**38.65**

Obrázek 4.3: Srovnání humanitních a přírodovědných oborů: procento termínových textových pozic. Nezávisle na celkovém počtu textových pozic v oboru (osa x) se procento termínů mezi textovými pozicemi v humanitních oborech pohybuje zhruba mezi 10 a 25 %, kdežto v oborech přírodovědných/technických zhruba mezi 20 a 35 %. Údaje jsou založeny na automatickém označení jednoslovných termínů.



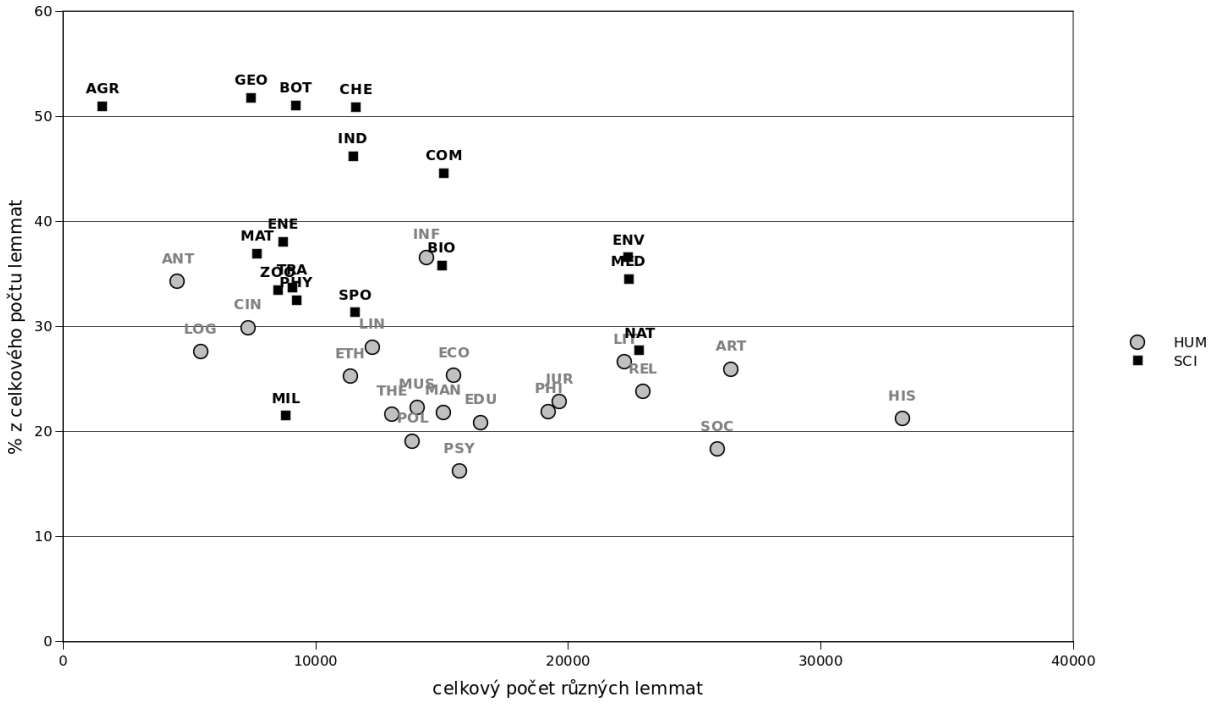
Obrázky 4.3 a 4.4 se zaměřují na srovnání humanitních a přírodovědných/technických oborů; je z nich zřejmé, že v přírodovědných/technických oborech je výrazně větší množství termínů v textech i mezi lemmaty.

V úvahu je třeba brát různé množství textů, které jsou k dispozici pro každou z disciplín. Procento termínů v textech není závislé na rozsahu dostupných textů, s velkou pravděpodobností se nebude zásadně lišit ani při zkoumání daleko rozsáhlejších dat. Počet jedinečných terminologických lemmat na rozsahu textů ale závisí, stejně jako je tomu u lemmat obecně. S přibývajícím počtem textových pozic přibývají další a další lemmata, resp. další a další termíny[7].

V relativně malém množství akademických textů dostupných pro každou disciplínu v korpusu SYN2010 nemůže být v žádném případě obsažena celá terminologie oboru, lze však předpokládat, že se zde vyskytne alespoň značná část terminologie centrální (nejčastější a nejdůležitější termíny). Okrajovější termíny spojené např. jen s určitým tématem se ve zkoumaných textech vyskytnou jen v závislosti na tématech těchto textů.

[7]Lemmata ani termíny přitom nepřibývají donekonečna - tomuto problému se věnoval Cvrček (2014).

Obrázek 4.4: Srovnání humanitních a přírodovědných oborů: procento termínových lemmat. I v tomto případě se ukazuje, že humanitní a přírodovědné/technické obory se odlišují, a to nezávisle na celkovém počtu jedinečných lemmat v jednotlivých disciplínách. Procento termínů mezi lemmaty se v humanitních oborech pohybuje s výjimkami mezi 15 a 30 %, kdežto v oborech přírodovědných/technických zhruba mezi 30 a 55 %. Údaje jsou založeny na automatickém označení jednoslovných termínů.



K automaticky vyhledaným termínům je tedy nutné přistupovat jen jako k součásti terminologie daného oboru[8]. Na základě předkládaných výsledků tedy nelze srovnávat velikost terminologie (ve smyslu úhrnu termínů jedné disciplíny), pouze procento terminologických lemmat ze všech lemmat, a to ještě s tou výhradou, že toto procento je závislé na velikosti dostupných oborových subkorpusů. V případě lemmat lze tedy sledovat pouze tendence, přičemž nejvhodnější by bylo srovnávat obory se zhruba stejným zastoupením (co do textových pozic) v korpusu SYN2010.

4.2 Jednoslovné termíny sdílené více obory

Parametry experimentu 8: Souvislosti mezi obory na základě sdílených jednoslovných termínů. V experimentu byly použity automaticky označené jednoslovné termíny v testovacích datech (z experimentu 7, podrobnosti viz výše). Jednotlivé obory byly porovnány na základě těchto automaticky vyhledaných termínů - zjišťovalo se, kolik termínů každá dvojice oborů sdílí. Subkorpusy oborů jsou různě rozsáhlé, proto je třeba chápat výsledky spíš orientačně.

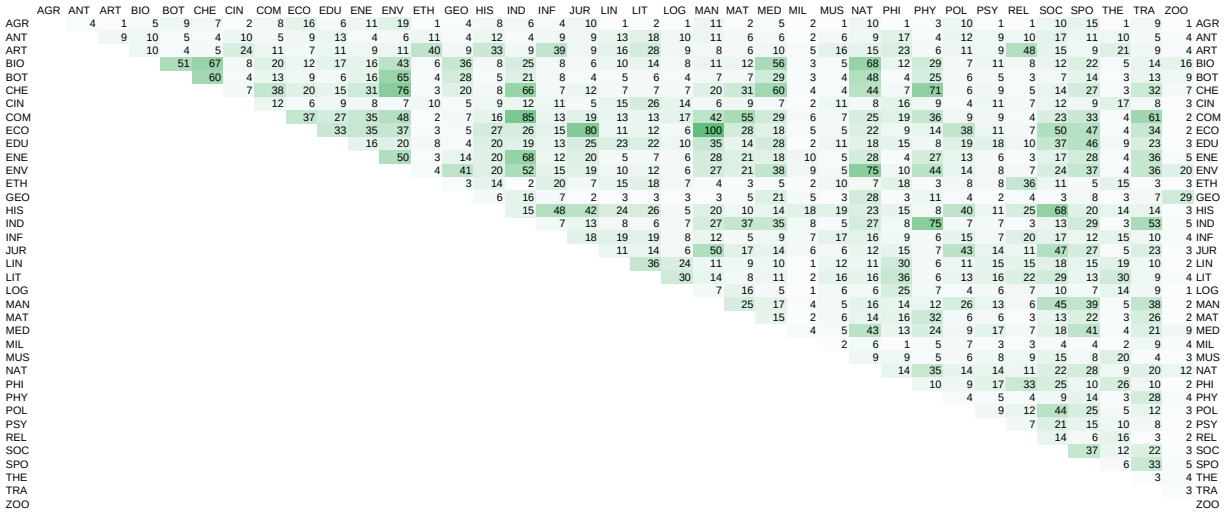
Na základě automaticky vyhledaných termínů lze zjišťovat, do jaké míry sdílejí různé obory termíny, přičemž větší počet společných termínů ukazuje na silnější vztah mezi disciplínami. Některé z těchto vztahů jsou intuitivně pociťovány (např. mezi technikou a informatikou, mezi botanikou a obecnou biologií), jinde může být vztah mezi obory především pro neodborníky zastřený (např. vztah mezi ekologií a chemií, sociologií a historií). Lze vysledovat i celé skupiny oborů, které vzájemně sdílejí termíny (obecná biologie, botanika, chemie, medicína), nebo třeba jeden obor, který ve velké míře přebírá termíny z mnoha dalších oborů (ekologie má velký počet termínů společný s botanikou, energetikou, geologií, chemií, technikou ad.).

Spolehlivost nalézání souvislostí mezi obory je jen relativní, a to především z toho důvodu, že množství textů pro různé disciplíny v korpusu SYN2010 se liší. Právě v tomto případě hraje velikost subkorpusu každé disciplíny značnou roli, zvláště v případě velmi málo zastoupených oborů[9], jako je zemědělství a antropologie (obory s nejmenším počtem textových pozic v korpusu SYN2010). Vztahy mezi obory mohou být navíc zvýrazněny nebo zamlženy výběrem jednotlivých akademických prací v korpusu SYN2010 (monografie z různých oborů, ale s podobným tématem velmi ovlivní množství společných termínů). Nelze se tedy spoléhat na absolutní hodnoty nebo procenta společných termínů, lze jen obecně konstatovat, že některé obory sdílejí (neobvykle) velké nebo malé množství termínů.

[8] Ucelenou terminologii oboru by bylo možné vyhledat jen v ideálním případě, že bychom měli k dispozici veškeré texty dané disciplíny.

[9] Malý počet textových pozic v datech oboru se výrazně odráží na velmi nízkém počtu sdílených termínů s jakýmkoli jiným oborem.

Obrázek 4.5: Zobrazení souvislostí mezi 37 obory dostupnými v SYN2010 na základě automaticky vyhledaných jednoslovných termínů společných jednotlivým dvojicím disciplín. Čím tmavší barva, tím více je mezi danými obory sdílených termínů. Hodnoty v obrázku jsou normalizovány (0-100), nevyjadřují tedy počet společných termínů.



Obrázek 4.5 ukazuje, mezi kterými obory jsou na základě dat z korpusu SYN2010 nejsilnější vztahy, a které z oborů sdílejí jen minimální počet automaticky vyhledaných termínů. Z důvodu větší srozumitelnosti je síla vztahů na škále mezi 1 a 100, kde 1 znamená nejslabší vztah (pouze jednotky sdílených termínů) a 100 je vůbec nejvyšší nalezená hodnota termínů sdílených dvěma obory (management a ekonomie, více než tisíc sdílených termínů). V oborech v korpusu SYN2010 byly za jednoslovné termíny označeny necelé dva miliony textových pozic z celkových 9 milionů, nejvyšší počet termínů společných dvěma disciplínám je zhruba 1600 (obory MAN a ECO, viz níže).

Nejsilnější vztahy (na základě společných termínů) jsou mezi následujícími obory: management a ekonomie, technika a informatika, právo a ekonomie.

Mezi dvojice s velmi malým množstvím sdílených termínů (desítky termínů) patří například[10] lingvistika a zoologie, management a filmová věda, historie a botanika, doprava a dějiny umění.

Pro představu je v následujícím přehledu zachyceno několik náhodně vybraných termínů, které jsou společné dvojici disciplín. Mezi prvními čtyřmi dvojicemi jsou takové obory, které sdílejí velký počet termínů, další čtyři dvojice patří mezi obory s velmi malým množstvím společných termínů.

[10]Záměrně zde nejsou zahrnuty obory s nejnižším počtem textových pozic, u nichž by výsledky byly zkreslené.

Velký počet sdílených termínů:

IND (průmysl) - COM (informatika): 1338 sdílených termínů

např. *algoritmus, amplituda, chybová, digitálně, elektromagnetickou, frekvence, gradient, impedance, kabel, kondenzátor, linearizace, napěťový, nastaven, parametr, převodník, protokolu, synchronní, tlačítka, tranzistor, vzorkuje*

JUR (právo) - ECO (ekonomie): 1255 sdílených termínů

např. *akcionář, celních, dlužník, faktury, hospodářský, komplementář, legislativa, likvidace, měnové, nabyvatel, nekalá, obchodněprávní, osoba, plnění, podniky, pravomoci, prodejem, sazebník, úprava, vlastník*

SOC (sociologie) - HIS (historie): 1070 sdílených termínů

např. *administrativy, agitační, byrokratizace, členů, demokracie, družstva, ekonomický, elit, hospodaření, kampaň, kultura, liberální, manifestace, mobilizace, orgány, propaganda, referenti, sjezd, výbor, znárodnění*

CHE (chemie) - BIO (biologie): 1045 sdílených termínů

např. *atomy, barviva, bakterie, bílkoviny, cukry, enzym, hydrolýza, koncentrace, metabolit, oxid, pH, rozpustný, sodík, termodynamický, živočišné, žlázy, tkáně, roztok, plastidy, morfologie*

Malý počet sdílených termínů:

TRA (doprava) - BIO (biologie): 213 sdílených termínů

např. *ekologické, ekosystém, energetický, kódy, odolnost, paměťových, reflexní, signály, transport, ventilace*

TRA (doprava) - LIN (lingvistika): 156 sdílených termínů

např. *komunikace, kvantifikátory, objektu, participant, referenční, sémantiky, singulárního, stavový, verbálních*

PHY (fyzika) - PHI (filozofie): 151 sdílených termínů

např. *aproximací, Aristotelés, derivace, eukleidovské, fyzika, makrosvěta, nekonečna, postulátů, teorém, vesmír*

LIN (lingvistika) - PHY (fyzika): 99 sdílených termínů

např. *částice, eliptický, frekvence, jaderných, konjugace, kvantitativní, nepřechodných, singularita, synchronním, valenční*

4.3 Rozbor automaticky vyhledaných jednoslovných termínů

Parametry experimentu 9: Vyhledání a analýza jednoslovných termínů v trénovacích datech. Experiment se soustředí na automatické vyhledání termínů v trénovacích datech z oborů COM, LIT, MED a SOC. Zaměřuje se na termíny s nejvyšší automaticky přidělenou terminologickou platností, na netermíny s nejnižší terminologickou platností, a dále na případy, kdy se ruční vyhledání liší od automatického. Materiál je v podobě nelemmatizované, s opakováním a bez slovních druhů. V experimentu byly použity čtyři vlastnosti: RFQdiscRFQcompar, ARF, RDist, a SDRD (viz kap. 5.1.3). Hranice mezi termíny a netermíny je v nástroji Weka defaultně nastavena na hodnotu 0,5.

4.3.1 Nejsilnější termíny

V tabulce 4.6 jsou uvedeny příklady nejsilnějších termínů vyhledaných v trénovacích datech ze čtyř oborů (celkem 8 tisíc textových pozic). V každém z těchto oborů, se mezi termíny s největší terminologickou platností zařazují jak slova srozumitelná jen pro odborníky, jako jsou *neasemblují, narativní, Trikuspidální, prekarizaci*, tak slova, kterým porozumí laik se středoškolským vzděláním, např. *RAM, literární, urogenitálního, sociologie*. Mezi nejsilnějšími termíny se ale objevují i slova známá nejširší veřejnosti (i když v obecnějším, neterminologickém významu), např. *adresa, vyprávění, dýchání, stát*.

Podívejme se na specifika nejsilnějších termínů každého ze zkoumaných oborů:

Mezi nejsilnějšími termíny v informatice (COM) jsou časté výrazně technické termíny (*byte, skripty, port, RAM*), ale i slova běžná v každodenní slovní zásobě užívané na pracovištích, jako jsou termíny *Windows, serveru, uživatel*. Mezi velmi silné termíny jsou zařazeny i zkratky pro fyzikální veličiny, např. *t* pro čas.

V literární vědě (LIT) jsou do této skupiny často zařazeny termíny sdílené s jinými obory (uměnovědy: *dílo, umění, autorské*, lingvistika: *textu, významy*). Mnohdy zde také nalezneme termíny ad hoc použité třeba jen v jednom díle, jako je *schizothymnosti, Kryptosémičnost*, které bývají málo srozumitelné nejen pro laiky, ale mnohdy i pro samotné odborníky. Na druhou stranu jsou mezi těmi nejvýraznějšími termíny v literární vědě i slova všeobecně známá široké veřejnosti: *básní, vyprávění, vypravěč*.

V lékařství (MED) patří mezi nejsilnější termíny vysoce specializované termíny vyskytující se často jen v tomto oboru a strukturou se odlišující od běžné české slovní zásoby (*allografty, athetoidními, Triskupidální, kadaverózní*), ale také termíny používané v neodborných textech v ne striktně terminologickém významu (*astmatu, dýchání, postižení*).

Tabulka 4.6: Příklady nejsilnějších automaticky označených termínů v oborech COM, LIT, MED a SOC (trénovací data). Příklady jsou vybírány z pozic s nejvyšší terminologickou platností: 1 až 0,97 (v oboru sociologie 1 až 0,87). Seznam všech automaticky vyhledaných termínů je v příloze D.

COM	LIT	MED	SOC
adresa	anticko-evropské	allografty	blahobytu
byte	autorské	astmatu	modernizace
data	básní	athetoidními	organizací
dobývání	čtenářových	autografty	péče
f	dílo	bovinního	politiky
frekvence	emitora	dýchacích	potřeb
hexadecimálním	fikční	dýchání	práva
HTML	fikčnost	kadaverózní	právněteoretických
instrukce	Kritikova	karcinoidním	prekarizaci
n	Kryptosémičnost	kolapsovými	rozhodování
neasemblují	literární	neurodermitis	služeb
podprogramu	narativního	postižení	socialismu
port	predeskribování	pupilární	sociálně
RAM	předtextové	sekretomotorické	sociálních
sebemodifikující	Styl	Stentless	sociálního
serveru	textu	stentované	sociologie
skripty	umění	tremor	stát
t	vypravěč	Trikuspidální	státem
uživatel	vyprávění	urogenitálního	trhu
Windows	významy	zbytněním	zabezpečení

Velmi silné termíny z oboru sociologie (SOC), který mimochodem sdílí velkou část slovní zásoby s politologií a ekonomií, se z větší části objevují i v publicistice: *politiky, práva, služeb, sociálních, stát, trhu*, a jsou tedy srozumitelné čtenářům publicistických textů. Některé ze silných termínů jsou v jiném (neterminologickém) významu používány v běžné slovní zásobě (*péče, rozhodování*).

4.3.2 Slovní druh automaticky vyhledaných termínů

Mezi termíny jsou téměř výhradně zařazena podstatná a přídavná jména, zcela výjimečně i sloveso (*neasemblují* COM, *rekonstruuje* LIT, *aktualizuje* LIT) nebo adverbium (*sociálně* SOC). V rámci trénovacích dat nebylo jako termín označeno slovo z žádného jiného slovního druhu[11] (o číslech v rámci termínů viz níže).

4.3.3 Čísla jako termíny

Číslům byla ve výzkumu věnovaná jen malá pozornost, ukázalo se však, že přinejmenším krátkou poznámku si zaslouží. Z trénovacích dat byla při data miningu všechna čísla vyřazena, a to kvůli své velké rozmanitosti (mohla by narušit trénovací proces). V testovacích datech ze všech akademických oborů však čísla (i v kombinaci s dalšími znaky) zůstala zachována a některá byla označena za termíny. Z nich je vybráno několik příkladů, které naznačují, že skutečně i mezi čísly (příp. mezi kombinacemi čísel a dalších znaků) lze nalézt termíny charakteristické pro danou disciplínu:

COM: *111.222.233.128* (IP adresa), *00001010* (zápis v dvojkové soustavě), *16bitový, 32bitový*

JUR: *140/1999, 167/2004* (čísla zákonů)

ECO: *3500Kč/měs, 1$/hod, 19069,-, 167/2004*

PHY: *-273°* (absolutní nula), *950°*

REL: *3,14* (2. Mojžíšova 3,14)

CHE: *100%ního* (roztoku), *250°*

[11] Více o slovních druzích v rámci termínů v kap. 1.2.2.2 a 4.4.

4.3.4 Nejsilnější netermíny

Mezi textové pozice, kterým byla automaticky přiřazena nejnižší hodnota terminologické platnosti (nula na ose 0 až 1), patří slova z téměř všech slovních druhů (výjimkou jsou jen citoslovce[12]). V tabulce 4.7 jsou uvedeny všechny netermíny z trénovacích dat s nulovou terminologickou platností.

4.3.5 Rozdíly mezi ručním a automatickým vyhledáváním

Mezi výsledky ručního a automatického vyhledávání termínů existují rozdíly, které lze obvykle charakterizovat jako chyby automatického zpracování. Ty jsou způsobené ve většině případů zvláštností slovního tvaru, např. velkým písmenem na začátku slova. Výjimečně dochází na základě automatického vyhledávání termínů k přehodnocení ručně označeného netermínu jako termínu (a naopak).

V tabulce 4.8 je přehled počtu chyb v automatickém (či výjimečně v ručním) označení termínů a netermínů v trénovacích datech (po 2000 pozicích ve čtyřech vybraných oborech).

Důvody rozdílného přiřazení hodnot termín či netermín při ručním a automatickém označování lze rozdělit do několika skupin. V mnoha případech by v lemmatizovaném materiálu k podobným chybám v automatickém zařazení nedošlo - proto je také automatické zařazování lemmat obvykle úspěšnější než u slovních tvarů (viz kap. 2.2.6).

Netermín je často mezi termíny zařazen (FP, false positive, viz 2.4):

- protože v jiném oboru (k němuž se konkrétní část akademického textu vztahuje) je daný slovní tvar termínem: *úvěr* (COM), *oděr* (z pneumatik) (MED)
- protože jde o vlastní jméno: *Mukařovský* (LIT), *Seebohm* (SOC)
- protože jde o jediný výskyt slovního tvaru v SYN2010: *vyklápějícím* (MED), *ujednocen* (SOC), *Nejplnějším* (SOC)
- protože jde o výraz neobvykle často používaný jedním autorem (v rámci jednoho nebo i více akademických textů): *nespolehlivostí* (LIT), *kombinace* (COM)
- protože jde o termín nerozpoznaný v rámci ručního vyhledávání (je nutná oprava): *tradiční* (SOC).

[12] Číslovky jsou zastoupené jen málo členy (*druhý, dvě, První*), k číslům viz kap. 4.3.3.

Tabulka 4.7: Nejsilnější automaticky označené netermíny z oborů COM, LIT, MED a SOC (trénovací data). V tabulce jsou všechna lemmata (případně tvary), kterým byla přiřazena hodnota terminologické platnosti 0. Konkrétní tvary (nikoli lemmata) jsou do seznamu zařazeny v následujících případech: substantiva, která jsou s vysokou pravděpodobností součástí víceslovných předložek (nelze zpětně zjistit s jistotou) - v tom případě je do seznamu zařazena i daná víceslovná předložka; adjektiva a adverbia ve stupňovaném tvaru; číslovka *dvě* (ženský rod) a substantivum *let* (lemma *léto* nedává v tomto kontextu smysl). Do slovních druhů jsou slova řazena podle Slovníku spisovné češtiny (4. vyd., 2005) s výjimkou částic a víceslovných předložek (Čermák, 2010, s. 175-180). Použitá metoda Bagging-PART.

N	A	P	Num	V	Adv	Prep	Konj	Part
část	celý	co	druhý	být	dostatečně	během	a	jen
činnost	další	jeho	dvě	lze	jak	do	aby	mnohem
době	hlavní	každý	První	měnit	kde	k	ale	například
funkce	jiný	který		mít	méně	mezi	či	nejen
konci	možný	některý		moci	předem	mimo	i	poněkud
let	nezbytný	on		muset	současně	na	jako	pouze
místo	nižší	se		obsahovat	spolu	na rozdíl od	li	právě
množství	podobný	svůj		patřit	stejně	na základě	nebo	především
možnost	případný	ten		platit	tak	o	než	spíše
období	různý	tento		sloužit	většinou	od	pak	také
oblasti	současný	žádný		souviset	více	pod	proto	zároveň
podmínka	stejný			tvořit	výše	podle	protože	zcela
podobě	velký			využít	vždy	před	však	zejména
pozornost	větší					při	že	
praxe	vnitřní					pro	by	
případě	základní					proti	což	
řadě	zvláštní					s		
rámci						směrem k		
rozdíl						u		
schopnost						v		
situace						včetně		
směrem						ve vztahu k		
souvislosti						v podobě		
stav						v podobě		
systém						v případě		
vztahu						v řadě		
základě						v rámci		
způsob						v souvislosti s		
						vzhledem k		
						z		
						za		

Tabulka 4.8: Nesprávně automaticky zařazené textové pozice u oborů COM, LIT, MED a SOC (2 tisíce textových pozic v každém oboru). Označení FP (false positive) znamená netermín nesprávně zařazený mezi termíny, značka FN (false negative) označuje termíny nesprávně zařazené mezi netermíny. Rozdíly mezi ručním a automatickým zařazováním (zjednodušeně chyby automatického zařazování) jsou pouze v řádu desítek textových pozic.

obor	**FP**	**FN**	**celkem nesprávně zařazených**
COM	19	46	65
LIT	58	31	89
MED	28	81	109
SOC	24	49	73

Termín je nejčastěji mezi netermíny zařazen v těchto případech (FN, false negative, viz 2.4):

- protože jde o neobvyklý tvar, který má celkově nízkou frekvenci, ale vyskytuje se i v akademických textech, i v publicistice nebo dokonce v beletrii (často jde o tvary s prvním velkým písmenem): *zápise* (COM), *Adresu* (COM), *Literát* (LIT), *spisovatelův* (LIT), *Postindustriální* (SOC), *novorozenecká* (MED)
- protože slovo je polysémní a ve srovnávacím korpusu má jiné, často obecnější významy: *třídu* (COM), *stránku* (COM), *stáhnout* (COM)
- protože daný termín se často vyskytuje v publicistických textech: *umělecká* (LIT), *lécích* (MED), *společenskou* (SOC), *psychologie* (SOC)
- protože dané slovo se vyskytuje v textech beletristických a publicistických v neterminologickém významu (často jako součást kolokací): *zrak* (MED), *sluch* (MED), *rodina* (SOC)
- protože skutečně jde o netermín (chyba v ručním značkování): *psaním* (LIT), *asociace* (MED), *těla* (SOC).

4.4 Rozbor automaticky vyhledaných víceslovných termínů

Parametry experimentu 10: Vyhledání a analýza víceslovných termínů v trénovacích datech. Experiment se soustředí na automatické vyhledání víceslovných termínů v trénovacích datech z oborů COM, LIT, MED a SOC. Zaměřuje se na termíny s nejvyšší automaticky přidělenou terminologickou platností, na netermíny s nejnižší terminologickou platností, a dále na případy, kdy se ruční vyhledání liší od automatického. Materiál je v podobě bigramů, bez lemmatizace, s opakováním a bez slovních druhů. V experimentu bylo použito šest vybraných vlastností: MWT:T1, MWT:T2, MWT:t-score, MWT:MI-score, MWT:Oblig a MWT:Prox (viz kap. 5.2.3). Hranice mezi termíny a netermíny je v nástroji Weka defaultně nastavena na hodnotu 0,5.

Akademické disciplíny mají svoje specifika ve vytváření a užívání víceslovných termínů - a můžou se od sebe lišit zásadním způsobem. Bylo by asi možné najít některé podobnosti mezi skupinami oborů s nějakým společným rysem, např. mezi uměnovědami oproti ekonomickým oborům, mezi teoretickými a praktickými obory nebo mezi obory humanitními oproti přírodovědným nebo technickým.

Díky rozsáhlým datům ze čtyř velmi odlišných disciplín, která jsou ručně i automaticky označkovaná, je možné zachytit alespoň některé obecnější rysy víceslovných termínů[13], jako je jejich rozsah, složení nebo funkce v textu, případně i výše zmiňované podobnosti či rozdíly mezi typy oborů.

V informatice se více než v jiných oborech setkáme v rámci víceslovných termínů se zkratkami, které vlastně také zastupují víceslovné termíny (viz níže), jako je *protokol HTTP* nebo *metodika CRISP-DM*. Literární věda formuje víceslovné termíny kratší než ostatní disciplíny, zcela převažují termíny dvojslovné. V lékařství je naopak největší výskyt mnohaslovných termínů, a zároveň jsou zde vůbec nejdelší termíny (v rámci zkoumaných textů až šestislovné, např. *aktuální ovlivnění autonomní regulace srdeční činnosti*). V sociologii je na rozdíl od ostatních tří oborů velmi malý počet jednoslovných a vysoký počet víceslovných termínů.

Většina termínů v textech akademických oborů je dvouslovná nebo trojslovná, delší termíny se vyskytují v menší míře[14]. V oboru LIT jsou termíny spíš jen dvouslovné, naopak v MED je dvouslovných termínů méně než v ostatních disciplínách a je zde tendence tvořit delší termíny.

Příklady čtyř- až šestislovných termínů: *dobývání znalostí z databází* (COM), *symbolické metody strojového učení* (COM), *kvartérní sektor národního hospodářství* (SOC), *srdečních*

[13]Tyto obecnější poznatky o termínech nelze aplikovat na všechny akademické obory bez výjimky. Texty v některých oborech, např. v chemii, můžou být natolik specifické, že pro ně nemusí platit žádná z obecných vlastností.

[14]Právě v délce víceslovných termínů se budou některé obory, jako je např. chemie, od ostatních disciplín velmi odlišovat

autonomních gangliových buněk (MED), *spektrální analýzy variability srdeční frekvence* (MED), *myelinizovaných nervových vláken periferních struktur ANS* (MED).

Naprostá většina víceslovných termínů (cca 95 %) obsahuje alespoň jeden termín a zhruba polovina dvouslovných termínů je dokonce složená ze dvou termínů (*datový byte* (COM), *umělecké dílo* (LIT), *endokarditida nativní* (MED), *tržní hospodářství* (SOC)). Výjimečně se vyskytují i víceslovné termíny tvořené pouze netermíny, jako je *umělá inteligence* (COM), *Umberto Eco*[15] (LIT) nebo *běžného života* (SOC), *soukromý život* (SOC) či *zdravý rozum* (SOC). V sociologii, jejíž terminologie je víc než u ostatních oborů propojena s neakademickými, především publicistickými texty (viz kap. 1.3.1.4 o terminologizaci obecných slov a determinologizaci termínů), je takových víceslovných termínů víc než v jiných oborech.

Nejčastější kombinací slovních druhů ve víceslovných termínech je kombinace adjektiva a substantiva nebo substantiva a substantiva a další kombinace podstatných a přídavných jmen. V naprosté většině případů má víceslovný termín v textech funkci substantiva (resp. jmenné fráze), výjimečně i adjektiva nebo i adverbia: *dálkově řízené (servopohony)* (COM), *sociálně vyspělých (států)* (SOC), *(natočený) koncem pánevním* (MED).

Příklady nejobvyklejších kombinací slovních druhů:

A-S: *webový prohlížeč* (COM), *tematický plán* (LIT), *vegetotropních farmak* (MED), *sociální stát* (SOC)

S-S: *registr procesoru* (COM), *původce textu* (LIT), *poruchy citlivosti* (MED), *flexibilizace práce* (SOC)

A-A-S: *vnější datovou pamětí* (COM), *cévní mozková příhoda* (MED), *nestátní sociální instituce* (SOC), *umělecké dílo literární* (LIT)

S-A-S: *testování klasifikačních znalostí* (COM), *ontologie literárního díla* (LIT), *regurgitace pulmonální chlopně* (MED), *stát sociálního zabezpečení* (SOC)

S-zkratka s funkcí S (nebo obrácené pořadí): *protokol HTTP* (COM), *IP síť* (COM), *poruchy ANS* (MED), *pH inkompatibilita* (MED)

Součástí víceslovných termínů bývají (v daleko menší míře) taky adverbia (***časově*** *kritické řízení* (COM)) nebo předložky: *počítače* ***se*** *zjednodušeným souborem instrukcí* (COM),

[15] V některých oborech jsou za víceslovné termíny označována i jména známých myslitelů, kteří se zasadili o rozvoj dané disciplíny. Mohou být považována za termíny určitého typu, už z toho důvodu, že méně známí autoři obvykle nebývají zmiňováni celým jménem.

Tabulka 4.9: Příklady nejsilnějších automaticky označených termínů v oborech COM, LIT, MED a SOC. Příklady jsou vybírány z bigramů s nejvyšší terminologickou platností (mezi hodnotami 1 a 0,9). Kompletní seznam automaticky vyhledaných termínů je k dispozici v příloze D.

COM	**LIT**	**MED**	**SOC**
terminálového programu	literárním textu	pulmonální chlopně	Stratifikace společnosti
HTML dokumenty	literární dílo	Echokardiografické vyšetření	sociálního tržního
IP adresu	narativního textu	medikamentózní léčbě	stát blahobytu
webový prohlížeč	umělecké dílo	aortální chlopeň	sociální služby
programové paměti	významového dění	mechanických chlopní	světového trhu
báze znalostí	fikčních světů	srdečních chlopní	sociálního státu
dobývání znalostí	predeskribování významů	ischemické choroby	tržního hospodářství
sériový port	postmodernistické dění	mozkové příhody	veřejných institucí

dobývání znalostí ***z*** *databází* (COM), *žlázy* ***s*** *vnitřní sekrecí* (MED). Možnost zapojení ostatních slovních druhů (snad jen s výjimkou citoslovcí) nelze vyloučit, ale zcela jistě jde o ojedinělé případy.

4.4.1 Nejsilnější víceslovné termíny a nejsilnější netermíny

Stejně jako jednoslovným, i víceslovným termínům byla automaticky přiřazena hodnota terminologické platnosti mezi hodnotami 0 a 1. V tabulce 4.9 jsou uvedeny příklady nejsilnějších víceslovných termínů (bigramy s nejvyšší terminologickou platností). Některé z nich jsou součástí delšího víceslovného termínu, jako např. *dobývání znalostí (z databáze)* (COM), *ischemické choroby (srdeční)* (MED), *sociálního tržního (hospodářství)* (SOC).

Nejsilnější netermíny většinou obsahují interpunkci (např. i závorky), předložky, zájmena, částice a tvary slovesa *být*. V bigramu s nejnižší terminologickou platností může jedním členem být dokonce i termín[16].

4.4.2 Vyhledání víceslovných termínů ze tří a více složek

Ač bylo původním záměrem označovat pouze termíny dvouslovné (bigramy), metoda automatického vyhledávání termínů TERMIT dokázala na základě natrénování označit i několik po sobě jdoucích bigramů, a tak vyhledat až šestislovné termíny. Dokonce i případy víceslovných termínů, jejichž součástí je předložka, např. *žlázy s vnitřní sekrecí*, byly v testovacích datech automaticky označeny za termíny (po sobě jdoucí bigramy *žlázy s, s vnitřní,*

[16]Příklady nejsilnějších netermínů obsahujících jeden termín: *prohlížeč s, báseň (, mozku je, než společností.*

Tabulka 4.10: Ukázka automaticky vyhledaných víceslovných termínů složených ze tří a více složek s terminologickou platností přidělenou každému z bigramů.

Bigram	**Term. platnost**
spektrální analýzy	0.86
analýzy variability	0.80
variability srdeční	0.94
srdeční frekvence	0.98
perineurální sítě	0.78
sítě argyrofilních	0.60
argyrofilních vláken	0.88
dobývání znalostí	0.99
znalostí z	0.86
z databází	0.66
metody strojového	0.93
strojového učení	0.99
politické preference	0.95
preference voličů	0.66
sociální politiky	0.99
politiky státu	0.89
hloubkovou strukturu	0.65
strukturu vyprávění	0.62
ontologie literárního	0.94
literárního díla	0.99

vnitřní sekrecí). V tab. 4.10 je několik příkladů takovýchto tří- až šestislovných termínů spolu s hodnotou terminologické platnosti každého bigramu daného termínu.

Automatické označení víceslovných termínů složených ze tří a více složek lze považovat za velký úspěch metody.

4.5 Shrnutí

Kapitola 4 je věnována automaticky vyhledaným jednoslovným a víceslovným termínům. Počet jednoslovných termínů v textech se pohybuje mezi 16 a 22 %, přičemž zhruba 60 % jednoslovných termínů stojí samostatně a 40 % je součástí víceslovných termínů.

Počet víceslovných termínů v textech je vždy nižší než počet jednoslovných. Zhruba 13 % textových pozic ze čtyř oborů je obsazeno víceslovnými termíny, přičemž ty jsou nejčastěji

složeny ze dvou či tří textových pozic. Výjimkou ale nejsou ani termíny čtyřslovné, v textu oboru lékařství byl nalezen i termín šestislovný. Počet víceslovných termínů (tedy nikoli obsazených textových pozic) se pohybuje mezi 50 až 70 termíny na tisíc textových pozic. Naprostá většina víceslovných termínů je tvořena alespoň z jednoho jednoslovného, celkově však jednoslovné termíny tvoří 75 % textových pozic obsazených víceslovnými termíny.

V humanitních vědách (vč. sociálněvědných) je výrazně menší množství termínů jednoslovných i víceslovných než v textech oborů technických či přírodovědných. Stejně tak je tomu i u počtu jednoslovných termínů ve slovní zásobě jednotlivých oborů.

Na základě automaticky vyhledaných termínů lze zjišťovat souvislosti či příbuznost mezi jednotlivými obory - čím více termínů dvě disciplíny sdílí, tím užší je mezi nimi vztah. Nejvíce sdílených termínů je mezi obory management a ekonomie, informatika a technika a právo a ekonomie. Za příklad málo těsných vztahů může posloužit příklad disciplín, jako je lingvistika a doprava či fyzika.

Z automaticky označených dat ze čtyř oborů (COM, LIT, MED a SOC) byly vybrány příklady termínů s nejvyšší automaticky přidělenou terminologickou platností, ať už jednoslovných (*byte, fikční, neurodermitis, péče*), nebo víceslovných (*HTML dokumenty, literární dílo, aortální chlopeň, sociální služby*). Rovněž byly uvedeny příklady netermínů s nejnižší automaticky přiřazenou terminologickou platností, mezi něž patří především funkční slova jako předložky a spojky, ale i slova ze všech ostatních slovních druhů kromě citoslovcí, které se v akademických textech vyskytují jen výjimečně. Mezi příklady z různých slovních druhů patří slova *část, stejný, který, druhý, patřit, dostatečně, na rozdíl od, což, poněkud*.

Mezi automaticky vyhledanými jednoslovnými termíny se vyskytují především substantiva a adjektiva, v menší míře pak slovesa a adverbia. Víceslovné termíny jsou nejčastěji dvou a tříslovné a mají obvykle funkci substantivní a bývají složeny z různých kombinací adjektiv a substantiv, v některých oborech i zkratek. Jedním z úspěchů metody TERMIT je vyhledání víceslovných termínů složených z více než dvou textových pozic. Nejdelším nalezeným termínem je termín šestislovný: *aktuální ovlivnění autonomní regulace srdeční činnosti*.

5 Vlastnosti termínů

Jedním z cílů této studie je získat nové poznatky o vlastnostech jednoslovných a víceslovných termínů. Následující kapitola je proto zaměřena na seřazení použitých vlastností podle jejich důležitosti pro proces automatického vyhledávání termínů.

Kapitola je rozdělena do dvou částí, věnovaných jednoslovným a víceslovným termínům. Každá část se kromě seřazení vlastností zabývá i sestavením vhodné kombinace menšího počtu rysů, na jejímž základě lze efektivně automaticky identifikovat termíny. Vytvoření takové kombinace má velký význam pro budoucí výzkumy automatického vyhledávání termínů.

5.1 Vlastnosti jednoslovných termínů

5.1.1 Seřazení vlastností dle důležitosti

Parametry experimentu 11: Seřazení vlastností jednoslovných termínů podle důležitosti. V experimentu bylo pomocí 13 metod z nástroje Weka, které jsou schopné ohodnocovat vlastnosti (mají funkci feature ranking nebo feature selection), seřazeny všechny používané vlastnosti jednoslovných termínů podle role, jakou hrají v automatickém vyhledávání termínů. Každá z dostupných metod, která má funkci feature ranking nebo feature selection, byla natrénována na 8 tisících textových pozic (trénovací data ALL). Materiál je ve formě nelemmatizované, s opakováním a bez rozlišení slovních druhů. Hranice mezi termíny a netermíny je v nástroji Weka defaultně nastavena na hodnotu 0,5.

S vytěžováním dat souvisí proces nazývaný feature ranking nebo feature selection (seřazení nebo výběr vlastností podle důležitosti). Při tomto procesu se hodnotí role jednotlivých vlastností v řešení dané úlohy; atributy se buď seřadí podle relevance pro daný úkol (feature ranking), nebo se vybere sada důležitých atributů (feature selection) (Witten, 2005).

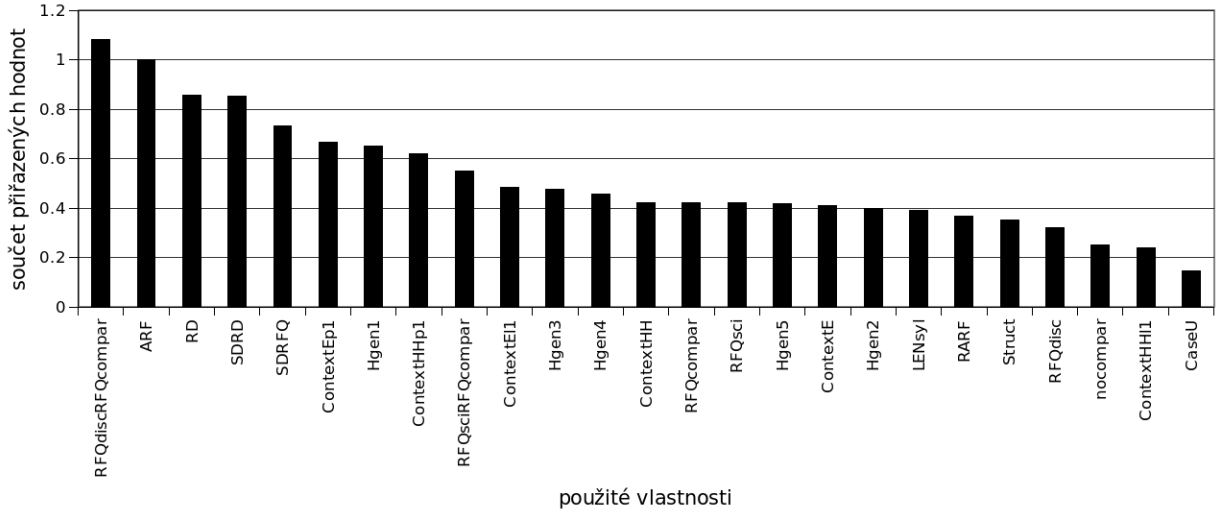
Funkce feature ranking nebo feature selection je součástí některých metod obsažených v data-miningových nástrojích, např. Weka (viz kap. 2.3.3). Tyto metody tedy poskytují příležitost seřadit statistické a další rysy jednotlivých textových pozic podle jejich důležitosti pro automatické vyhledávání termínů, případně vybrat sadu nejdůležitějších rysů (tj. takovou, která umožní vysokou úspěšnost metody).

Zjištění relevance jednotlivých vlastností má zásadní význam ze tří důvodů: 1. rysy, které při automatickém vyhledávání hrají velkou roli, můžeme považovat za charakteristické pro termín (mohou tudíž být zahrnuty do popisu termínu), 2. zjištění, které ze zkoumaných vlastností jsou skutečně relevantní, má velký význam pro další výzkumy v oblasti automatického vyhledávání termínů a 3. další experimenty lze provádět s daleko menším počtem rysů, tedy s menšími nároky na technické zpracování.

Tabulka 5.1: Feature ranking (seřazení dle důležitosti) všech použitých vlastností. Podstatné je zvláště pořadí a pak i rozdíl mezi jednotlivými hodnotami, nikoli sama výsledná hodnota (ta je pouze součtem hodnot přiřazených jednotlivými metodami v rámci nástroje Weka).

Pořadí	Atribut	Součet hodnot
1	RFQdiscRFQcompar	1.08
2	ARF	1.00
3	RDist	0.86
4	SDRD	0.85
5	SDRFQ	0.73
6	KontextEp1	0.67
7	Hgen1	0.65
8	KontextHHp1	0.62
9	RFQsciRFQcompar	0.55
10	KontextEl1	0.48
11	Hgen3	0.48
12	Hgen4	0.46
13	KontextHH	0.42
14	RFQcompar	0.42
15	RFQsci	0.42
16	Hgen5	0.42
17	KontextE	0.41
18	Hgen2	0.40
19	LENsyl	0.39
20	RARF	0.37
21	Struct	0.35
22	RFQdisc	0.32
23	NoCompar	0.25
24	KontextHHl1	0.24
25	CaseU	0.15

Obrázek 5.1: Feature ranking (seřazení dle důležitosti) všech použitých vlastností (viz tab. 5.1). Podstatné je pořadí, nikoli hodnota na ose y (ta je pouze součtem hodnot přiřazených jednotlivými metodami v rámci nástroje Weka).



Následující pořadí vlastností zahrnutých do předkládaného výzkumu (kap. 2.2.4, tabulka 2.4) je založeno na všech třinácti metodách v nástroji Weka, které buď seřazují atributy podle důležitosti (feature ranking), nebo z nich vybírají jen určitou účinnou sestavu (feature selection). Vyhodnocení proběhlo tak, že každému rysu byla připsána normalizovaná hodnota podle toho, na jaké místo byl zařazen, nebo podle toho, zda byl vybrán do sady důležitých rysů a kolik dalších rysů tato sestava obsahuje. Všechny hodnoty byly následně sečteny a vlastnosti byly seřazeny od té nejdůležitější až po nejméně důležitou.

Z uvedené tabulky 5.1 (resp. obrázku 5.1) je možné určit, jak je který rys ve srovnání s ostatními vlastnostmi důležitý pro proces automatického vyhledávání termínů. Zároveň je na jejím základě (podle pořadí) možno vytvořit menší sadu atributů, která by byla prakticky využitelná v dalších experimentech i v dalších výzkumech terminologie a ATR.

Podle feature rankingu založeném na data-miningových metodách v nástroji Weka jsou čtyři nejdůležitější rysy pro automatické vyhledávání termínů[1]:

1. poměr relativní frekvence v určité disciplíně ku relativní frekvenci ve srovnávacím korpusu (**RFQdiscRFQcompar**): čím je tento poměr vyšší, tím častěji se slovo vyskytuje v disciplíně oproti srovnávacímu korpusu, v extrémním případě se slovo ve

[1]K interpretaci vlivu jednotlivých atributů na terminologickou platnost viz i kap. 5.1.2, zvláště tab. 5.2.

srovnávacím korpusu neobjevuje vůbec. U takových slov je daleko vyšší pravděpodobnost, že půjde o termíny.

2. průměrná redukovaná četnost (pravidelnost rozmístění v korpusu) (**ARF**): čím je ARF vyšší, tím je slovo frekventovanější a rovnoměrněji rozloženo v korpusu (frekvence a rozloženost jsou zde velmi těsně propojeny, hodnota ARF vyjadřuje vždy obojí zároveň). Slova s nižší hodnotou ARF jsou s větší pravděpodobností termíny, protože v korpusu jsou rozložena nepravidelně[2], ve shlucích, tedy v jednotlivých disciplínách či jen v několika textech.

3. relativní distribuce textové pozice v akademických disciplínách (**RDist**): čím je relativní distribuce slova v oborech nižší, tím větší je pravděpodobnost, že jde o termín. Termíny jsou často omezeny na jeden nebo několik oborů (často jde o obory příbuzné), slova vyskytující se v mnoha oborech jsou buď slova běžná i v neakademických textech, nebo může jít o oborově nespecifické termíny (viz 1.3.1.2).

4. směrodatná odchylka od relativní vzdálenosti (**SDRD**): čím je tato směrodatná odchylka vyšší, tím nepravidelnější jsou rozestupy mezi jednotlivými výskyty slova a tím větší je i pravděpodobnost, že dané slovo je termínem. Jedná se o rys, který se v určitém ohledu podobá rysu ARF (průměrná redukovaná četnost), zásadní rozdíl je v tom, že nezohledňuje frekvenci zkoumaného slova[3].

5.1.2 Korelace vlastností

Parametry experimentu 12: Korelace vlastností jednoslovných termínů. V experimentu byla pomocí dataminingového nástroje RapidMiner vytvořena korelační matice všech rysů používaných ve výzkumu, vč. ručně přiřazovaného rysu „termín“. Jako materiál jsou použita sloučená trénovací data (ALL). Materiál je v podobě nelemmatizované, s opakováním a bez slovních druhů. Hranice mezi termíny a netermíny je v nástroji Weka defaultně nastavena na hodnotu 0,5.

Pro spolehlivou interpretaci výsledků feature rankingu je zapotřebí informace o tom, jakým způsobem spolu korelují jednotlivé atributy[4]. Korelace je jedním z nástrojů statistického popisu: Dvě proměnné spolu korelují, pokud se jejich hodnoty mění společně v určitých kvantifikovatelných krocích (Volín, 2007, s. 185). Zjišťuje se síla korelace (slabá nebo silná

[2]Termíny zároveň mají obvykle nižší frekvenci v korpusu ve srovnání např. s gramatickými slovy (frekvence je důležitou součástí hodnoty ARF).

[3]Z nízké korelace obou atributů (viz nízkou hodnotu v obrázku 5.2) je zřejmé, že oba rysy jsou si skutečně podobné jen okrajově.

[4]Pro měření korelace v experimentu 12 byla použita speciálně upravená data - z trénovacích dat byla navíc odstraněna interpunkce včetně závorek, uvozovek a pomlček, aby byl zredukován vliv těchto znamínek (z nichž některé jsou výrazně frekventované) na konečný výsledek. Interpunkční znamínka např. zkreslovala informace o rysech, na nichž se skrytě, ale výrazně podílí frekvence.

korelace), a také kladná či záporná hodnota[5]. Zásadní je si uvědomit, že vlastnosti, které spolu korelují, nemusejí jedna druhou způsobovat.

Korelace je podobná nejjednodušším statistickým data-miningovým metodám. Komplexnější data-miningové metody mají oproti korelaci lepší výsledky mj. proto, že vyhledávají různé, i daleko složitější (např. vícestupňové nebo vícerozměrné) vztahy v datech. Výsledky data-miningových nástrojů proto nelze nahradit hledáním korelací, ale jak už bylo zmíněno výše, korelace mezi jednotlivými atributy může být užitečným, a v našem případě i nezbytným, doplněním informací potřebných k interpretaci výsledků.

5.1.2.1 Korelace vlastností termínu a vysoké terminologické platnosti

Znázornění korelací mezi jednotlivými atributy může mít podobu korelační matice (obrázek 5.2), z níž je zřejmé, zda korelace mezi dvěma konkrétními rysy je kladná, či záporná, a jak vysokou má hodnotu (jak je silná či slabá).

Obrázek 5.2 obsahuje mj. i zásadní informaci o korelaci různých vlastností s atributem „termín“ (daná textová pozice, obv. slovo) je automaticky označena jako termín)[6]. Z pouhého seřazení rysů podle důležitosti (feature ranking, viz kap. 5.1.1) totiž na rozdíl od korelací nevyplývá, který z nich má pozitivní vliv na označení textové pozice za termín, a které atributy působí v tomto ohledu negativně (tedy čím je vyšší hodnota atributu, tím menší je pravděpodobnost zařazení textové pozice mezi termíny)[7]. Hodnoty korelace jednotlivých vlastností s atributem „termín“ jsou uvedeny také v tab. 5.2, přičemž nejsilnější korelace jsou zvýrazněny.

RFQdiscRFQcompar a termín (pozitivní korelace)

> Čím je poměr relativní frekvence slova v oboru ku relativní frekvenci ve srovnávacím korpusu vyšší, tedy čím častěji se slovo vyskytuje v disciplíně oproti srovnávacímu korpusu, tím vyšší je pravděpodobnost, že dané slovo je termínem.

[5]Kladná korelace: čím vyšší je hodnota A, tím vyšší je hodnota B; záporná korelace: čím vyšší je hodnota A, tím nižší je hodnota B.

[6]Korelace mezi daným rysem a pravděpodobností, že textová pozice je termínem: Pozitivní korelace znamená, že čím je hodnota rysu vyšší, tím vyšší je pravděpodobnost, že jde o termín, a naopak. Při negativní korelaci platí, že čím je hodnota rysu vyšší, tím nižší je pravděpodobnost, že daná textová pozice je termínem.

[7]Tatáž data je sice možné získat z průběhu experimentů, ale jde o poměrně náročnou práci (některé metody např. poskytnou soupis všech pravidel použitých k automatickému vyhledání termínů, ale takových pravidel je obvykle několik desítek, mnohdy ale jde i o stovky pravidel, která by bylo třeba ručně analyzovat).

Tabulka 5.2: Korelace zkoumaných vlastností s automaticky přiřazeným rysem „termín“ (řazeno dle pořadí ve feature rankingu, viz 5.1).

Pořadí v FR	Atribut	Korelace s termínem
1	**RFQdiscRFQcompar**	**0.60**
2	**ARF**	**-0.49**
3	**RDist**	**-0.55**
4	SDRD	0.37
5	SDRFQ	0.30
6	**KontextEp1**	**-0.46**
7	Hgen1	-0.25
8	KontextHHp1	0.18
9	**RFQsciRFQcompar**	**0.43**
10	KontextEl1	-0.22
11	Hgen3	-0.35
12	Hgen4	-0.40
13	KontextHH	-0.06
14	RFQcompar	-0.23
15	RFQsci	-0.23
16	**Hgen5**	**-0.42**
17	KontextE	-0.36
18	Hgen2	-0.27
19	LENsyl	0.34
20	RARF	0.22
21	Struct	-0.28
22	RFQdisc	-0.21
23	NoCompar	0.24
24	KontextHHl1	-0.07
25	CaseU	0.04

RDist a termín (negativní korelace)

Čím menší je relativní distribuce slova v oborech, tedy v čím menším počtu oborů se slovo vyskytuje, tím vyšší je pravděpodobnost, že se jedná o termín.

ARF a termín (negativní korelace)

Čím méně pravidelné je rozložení slova v korpusu (a také čím je slovo méně frekventované), tím větší je pravděpodobnost, že jde o termín.

KontextEp1 a termín (negativní korelace)

Čím nižší je entropie prvního pravého kontextu slova, tedy čím vyšší je míra uspořádanosti v pravém kontextu, tím vyšší je pravděpodobnost, že toto slovo je termín[8].

RFQsciRFQcompar a termín (pozitivní korelace)

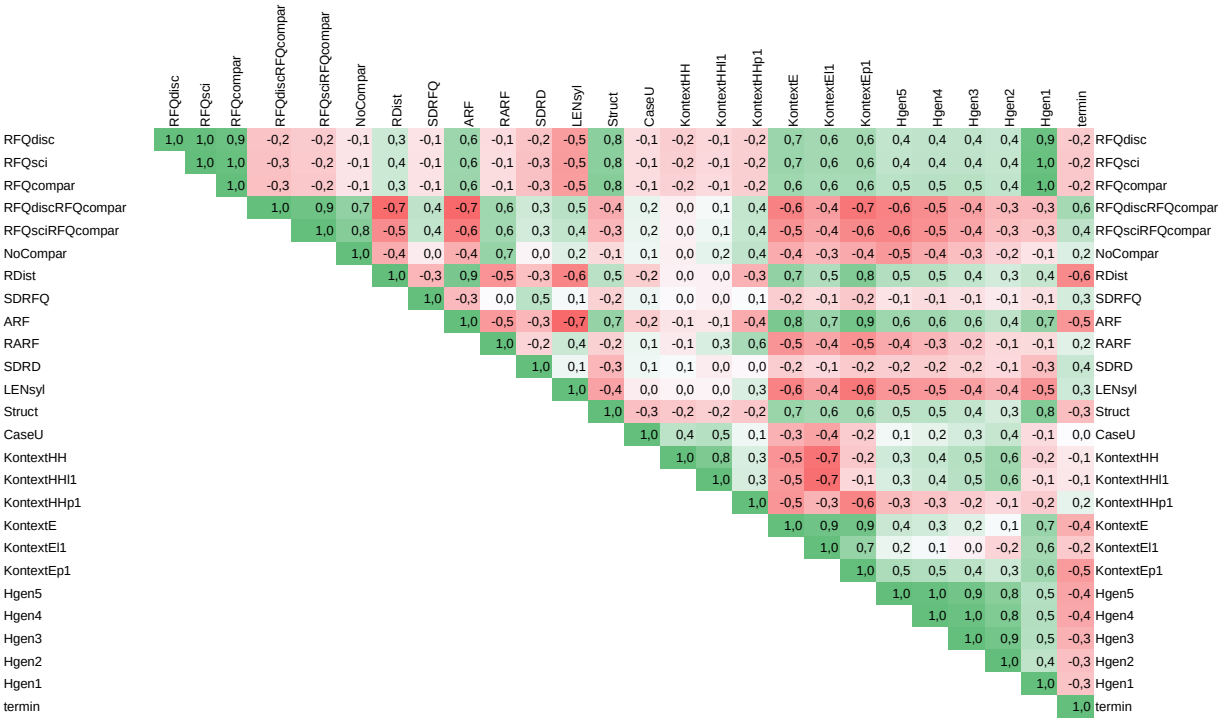
Čím vyšší je poměr relativní frekvence slova v akademickém subkorpusu SCI ku relativní frekvenci ve srovnávacím korpusu, tím větší je pravděpodobnost, že jde o termín. Oproti prvnímu rysu (RFQdiscRFQcompar) mohou taková slova být termínem, nebo patřit mezi oborově nespecifické termíny (viz kap. 1.3.1.2).

Hgen5 a termín (negativní korelace)

Čím vyšší je vážený průměr relativních frekvencí předcházejícího kontextu (5 slov předcházejících zkoumané textové pozici), tím menší je pravděpodobnost, že textová pozice je termínem. Velmi zjednodušeně lze říct, že termínu předcházejí spíš frekventovanější slova.

[8]Korelace obou rysů je vzájemná a závislost jednoho na druhém jde oběma směry. Zjednodušeně a obecněji lze říct, že slovo si vyžaduje kontext a kontext determinuje výběr slova. O jednotlivých kontextových rysech a o roli kontextu viz Cvrček (2013).

Obrázek 5.2: Korelační matice znázorňující korelace mezi jednotlivými zkoumanými rysy, vč. ručně přiřazeného rysu „termín“. Čím tmavší je barva, tím silnější je korelace mezi dvěma rysy, červená barva znamená kladou korelaci a zelená zápornou.



5.1.2.2 Korelace vlastností jednoslovných termínů

Obrázek 5.2 znázorňuje i korelace různých vlastností jednoslovných termínů mezi sebou. Následující část obsahuje komentáře k jednotlivým zajímavým místům korelační tabulky[9]:

Korelace RFQsciRFQcompar a RFQdiscRFQcompar

Subkorpusy jednotlivých disciplín jsou podmnožinami subkorpusu akademických textů. Pokud je tedy relativní frekvence slova v jednom konkrétním oboru vyšší oproti srovnávacímu korpusu, je obvykle vyšší i v celém subkorpusu akademických textů. Jde jak o termíny, tak i o oborově nespecifické termíny (viz 1.3.1.2) (např. *hypotéza* apod.). Pokud se slovo častěji vyskytuje ve srovnávacím korpusu oproti subkorpusu jedné akademické disciplíny (RFQsciRFQcompar má nižší hodnotu), často i v celkovém akademickém subkorpusu (SCI) budou jeho výskyty méně časté (např. slova jako *furt, bábovka* nebo *sáňkovat* se nebudou vyskytovat v akademickém subkorpusu SCI a tudíž ani v jediné disciplíně).

Negativní korelace RFQsciRFQcompar a ARF

Čím nižší je poměr relativní frekvence slova v akademických textech ku relativní frekvenci ve srovnávacím korpusu, tím rovnoměrnější je rozložení slova v korpusu. Rovnoměrně rozložená slova, jako jsou běžné předložky a spojky nebo třeba sloveso *být*, se s podobnou frekvencí vyskytují ve všech textech, akademických i neakademických. Naopak slova s vyšším poměrem relativní frekvence v oboru ku relativní frekvenci ve srovnávacím korpusu (jako např. *přeučení, nivelizace, chlopenní*) jsou v korpusu (např. SYN2010) obvykle rozložena nerovnoměrně, vyskytují se jen ve shlucích, v jedné nebo několika disciplínách, nebo dokonce jen v jednom nebo několika textech.

Negativní korelace RDist a RFQsciRFQcompar

Čím nižší je relativní distribuce slova v jednotlivých oborech, tím vyšší je poměr relativní frekvence slova v akademických textech ku relativní frekvenci ve srovnávacím korpusu. Jde o slova typu *gastrointestinální, juxtapozice, servopohon* apod., omezená na jeden nebo málo oborů - taková slova se v neakademických textech objevují jen výjimečně.

[9]Podrobná analýza a interpretace korelací rysů by vyžadovala rozsáhlejší samostatnou práci, zde pouze nabízím dílčí postřehy, abych naznačila, jak složité jsou vztahy mezi jednotlivými zkoumanými vlastnostmi.

Negativní korelace RFQdicsRFQcompar s KontextEp1 a dalšími kontextovými atributy

Čím vyšší je poměr relativní frekvence slova v akademických textech ku relativní frekvenci ve srovnávacím korpusu, tím nižší je entropie (neboli tím vyšší je míra uspořádanosti) prvního pravého kontextu slova. Lze to vysvětlit dvěma způsoby: 1. Slova vyskytující se mnohem víc v oboru oproti srovnávacímu korpusu často patří mezi termíny. Jednoslovné termíny bývají součástí víceslovných termínů - tím se zúží možnosti kombinovatelnosti slova a zvýší se míra uspořádanosti kontextu. 2. Ve skutečnosti však jde spíš o společnou skrytou korelaci s frekvencí. Velmi frekventovaná slova, např. běžné spojky nebo předložky, mají obvykle daleko neuspořádanější kontext už jen na základě své vysoké frekvence (slovo s frekvencí 20 může mít nanejvýš 20 různých kontextů na jedné z bezprostředních pozic, slovo s frekvencí 10 tisíc samozřejmě nesrovnatelně víc).

Korelace ARF s KontextEp1 a dalšími kontextovými atributy

Čím pravidelnější je rozložení slova v korpusu (SYN2010), tím vyšší je entropie jeho prvního pravého kontextu. Můžeme to vysvětlit dvěma způsoby (stejně jako v předcházejícím případě): 1. Slova s málo pravidelným rozložením v korpusu často patří mezi termíny, a jako takové bývají často součástí víceslovných termínů, čímž se snižuje kombinovatelnost takových slov a entropie kontextu. 2. Ve skutečnosti ale jde spíš o společnou skrytou korelaci s frekvencí. ARF je rys, který obsahuje velmi silnou informaci o frekvenci, čím vyšší ARF, tím vyšší i frekvence daného slova. Velmi frekventovaná slova mají obvykle daleko rozrůzněnější kontext.

Korelace RDist a ARF

Slova vyskytující se ve více disciplínách v subkorpusu akademických textů jsou obvykle rovnoměrněji rozložena v celém korpusu (SYN2010); už omezení na jedinou disciplínu při velmi nízké relativní distribuci negativně ovlivňuje celkové ARF. Slova, která se vyskytují jen v jednom oboru, se málokdy vyskytují v beletristických nebo publicistických textech, tudíž jejich rozložení v korpusu bude méně rovnoměrné.

Negativní korelace ARF a LENsyl

Čím je slovo kratší, tím pravidelněji bývá rozloženo v korpusu (např. SYN2010), a naopak. Velký vliv na tuto korelaci mají nejkratší slova, neslabičná nebo jednoslabičná (většinou spojky nebo předložky), která jsou v korpusu poměrně pravidelně

rozložená. Naopak velmi dlouhá slova (často cizího původu), která jsou termíny (např. *gastrointestinální*, *antiesencialismus*) jsou obvykle rozložena málo pravidelně[10].

Korelace ARF a Struct

Čím vyšší je ARF slova, tím je vyšší pravděpodobnost, že jeho struktura bude obvyklejší, pravidelnější - struktura slova se totiž vypočítává z korpusových dat, takže nejčastější slova se na hodnocení obvyklosti/pravidelnosti významně podílejí. Slova se strukturou neobvyklou (s písmeny g, f, a s kombinacemi jako r+i, e+u apod.) jsou obvykle přejatá z nějakého cizího jazyka a v korpusu (např. SYN2010) rozložena méně pravidelně, často právě v akademických textech.

Slabá korelace ARF a RARF

RARF (relativní průměrná redukovaná četnost) je rys odvozený od ARF (redukovaná četnost). Přesto spolu tyto rysy nekorelují příliš silně - v RARF je totiž zcela potlačená informace o frekvenci, která v ARF hraje velkou roli. RARF tedy neroste s frekvencí, jak je tomu u ARF, ale pouze s pravidelností rozložení slova v korpusu.

Korelace RDist a KontextEp1

Čím nižší je relativní distribuce v oborech (např. pokud se slovo vyskytuje jen v jednom nebo dvou oborech), tím nižší je entropie pravého kontextu slova. Nízká entropie kontextu znamená, že je kontext více uspořádaný. Lze to vysvětlit dvěma způsoby: 1. Slova nízkou relativní distribucí v disciplínách bývají často termíny, a jako takové bývají často součástí víceslovných termínů, čímž se snižuje kombinovatelnost takových slov a entropie kontextu. 2. Ve skutečnosti ale jde spíš o společnou skrytou korelaci s frekvencí. Slova s vysokou hodnotou RDist patří často mezi slova poměrně frekventovaná. Hodně frekventovaná slova mívají málo uspořádaný kontext oproti slovům málo frekventovaným (viz výše).

Korelace KontextEp1 a LENsyl a Struct

Čím delší slovo a čím složitější je jeho struktura, tím nižší je entropie jeho prvního pravého kontextu. Stejně jako v případě korelace entropie s RFQdiscRFQcompar a s ARF, i zde jde pravděpodobně o skrytou korelaci s frekvencí: delší a složitější

[10]Pro srovnání: hodnota ARF předložky *k* v SYN2010 je 273 tisíc, hodnota ARF termínu *gastrointestinální* je pouze 6.

slova mají spíš nižší frekvenci (oproti např. běžným spojkám a předložkám); slova s nižší frekvencí mají menší možnosti kombinovatelnosti, a tudíž mají i vyšší míru uspořádanosti kontextu (v tomto případě bezprostředního pravého kontextu).

5.1.3 Nejvhodnější kombinace vlastností pro ATR

Parametry experimentu 13: Nejvhodnější kombinace vlastností jednoslovných termínů. Experiment je zaměřen na nalezení vhodné kombinace vlastností sloužící k automatickému vyhledávání jednoslovných termínů (viz experimenty v kap. 3). Vhodnost kombinací se hodnotí na základě úspěšnosti metod PART, JRip, J48graft a Bagging-PART (z nástroje Weka) na sloučených trénovacích datech (ALL). Materiál je v podobě nelemmatizované, s opakováním a bez slovních druhů. Hranice mezi termíny a netermíny je v nástroji Weka defaultně nastavena na hodnotu 0,5.

Pro další výzkum v oblasti automatického vyhledávání termínů (ATR) by bylo velmi výhodné nabídnout nejvhodnější (tzn. efektivní) kombinaci atributů, na jejímž základě lze vyhledat velké množství termínů v textech. Stejně jako v některých předchozích experimentech, i zde je třeba vzít v úvahu dva vzájemně protichůdné požadavky: 1. co největší úspěšnost vyhledávání a 2. co nejjednodušší příprava a průběh experimentů (v tomto případě co nejmenší počet potřebných rysů)[11].

Prvním krokem pro nalezení takové efektivní kombinace je seřazení atributů podle důležitosti (feature ranking, kap. 5.1.1), následuje vyhodnocení úspěšnosti na datech s různým počtem rysů. V neposlední řadě je třeba sledovat i korelaci daného rysu s termínem, viz tab. 5.2.

V tomto experimentu je srovnávána úspěšnost čtyř metod (PART, JRip, J48graft a Bagging-PART) z data-miningového nástroje Weka při použití různých kombinací atributů podle tabulky 5.3 (prvních deset položek z tabulky 5.2 - feature ranking). K prvnímu, nejdůležitějšímu atributu se postupně přidávají další rysy, takže vznikne 10 kombinací (1, 1-2, 1-3 až 1-10). Jedenáctou kombinaci (uvedenou pro srovnání) tvoří všech 22 vlastností používaných pro vyhledávání jednoslovných termínů.

Obrázek 5.3 nabízí přehled toho, jak výrazně nebo málo výrazně mění přidávání a ubírání jednotlivých rysů úspěšnost experimentů. Rozdíl mezi výsledky se všemi a s deseti rysy je jen velmi malý. Vhodná by mohla být např. kombinace rysů 1-9 (vysoká úspěšnost u všech metod), ale její nevýhodou je vysoký počet rysů. Skutečný pokles úspěšnosti začíná u kombinace 1-3, větší skok je u následující kombinace (1-2). Nejvýhodnější se zdá být kombinace 1-4, kde ještě nezačíná pokles a zároveň jde o malý počet rysů.

Tabulka 5.4, která ukazuje průměrnou hodnotu míry *accuracy* ze 4 použitých metod, uka-

[11]Pokud by v dalších experimentech postačilo pro každou textovou pozici vypočítat hodnoty např. pěti atributů oproti patnácti atributům, znamenalo by to výrazné zjednodušení práce.

Tabulka 5.3: Přehled prvních deseti rysů vyhodnocených jako nejdůležitější feature rankingem (údaje převzaty z tab. 5.2).

Pořadí	Atribut
1	RFQdiscRFQcompar
2	ARF
3	RDist
4	SDRD
5	SDRFQ
6	KontextEp1
7	Hgen1
8	KontextHHp1
9	RFQsciRFQcompar
10	KontextEl1

zuje, že hodnota *accuracy* se nijak výrazně nezvyšuje při doplňování rysů 5 až 10 (nebo se dokonce snižuje). Ideální kombinací, tzn. nejjednodušší úspěšnou, je kombinace rysů 1 až 4.

5.2 Vlastnosti víceslovných termínů

5.2.1 Seřazení vlastnostní dle důležitosti

Parametry experimentu 14: Seřazení vlastností víceslovných termínů podle důležitosti. V experimentu byly pomocí metod z nástroje Weka, které jsou schopné ohodnocovat vlastnosti (feature ranking a feature selection), seřazeny všechny vlastnosti pro víceslovné termíny (označené MWT) podle důležitosti pro proces automatického vyhledávání. Každá z dostupných metod, která má funkci feature ranking nebo feature selection, byla natrénována na trénovací data (pouze ALL). Materiál je ve formě bigramů, bez lemmatizace, s opakováním a bez rozlišení slovních druhů. Hranice mezi termíny a netermíny je v nástroji Weka defaultně nastavena na hodnotu 0,5.

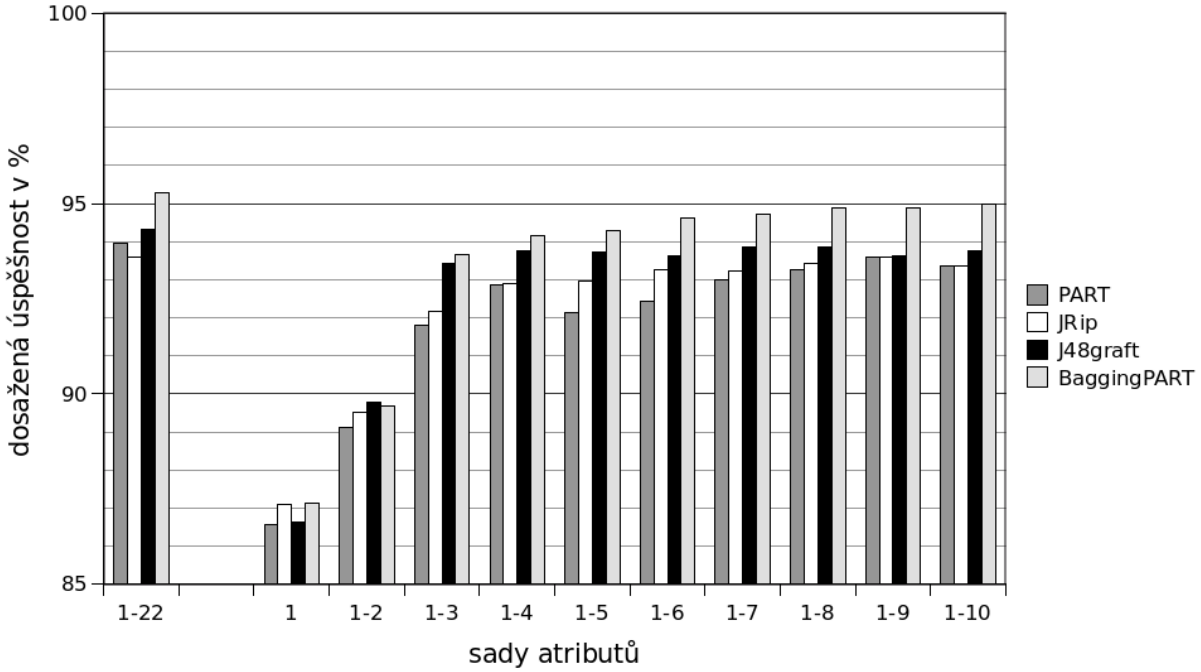
S vytěžováním dat souvisí proces nazývaný feature ranking nebo feature selection (seřazení nebo vlastností podle důležitosti). Při tomto procesu se hodnotí role jednotlivých vlastností[12] v řešení dané úlohy - rysy jsou seřazeny podle relevance pro daný úkol. Podle feature rankingu provedeného v data-miningovém nástroji Weka patří mezi nejdůležitější rysy při vyhledávání víceslovných termínů lexikální asociační míry t-score a MI-score, vysoká terminologická platnost prvního a druhého slova v bigramu[13] a kontextové vlastnosti obligatornost a proximita[14].

[12]Vlastnosti používané pro vyhledávání víceslovných termínů jsou podrobněji popsány v kap. 2.2.4 (vlastnosti s označením MWT).

[13]Při vyhledávání víceslovných termínů se nehodnotí jednotlivé textové pozice, ale bigramy složené ze dvou po sobě jdoucích textových pozic (bez interpunkce).

[14]Na rozdíl od jednoslovných termínů je při vyhledávání víceslovných použita sada šesti rysů (u jedno-

Obrázek 5.3: Srovnání úspěšnosti označování termínů s různými kombinacemi atributů. Atributy jsou řazeny podle výsledků feature rankingu (viz tab. 5.3, příp. kap. 5.1.1). Úspěšnost (hodnota míry *accuracy*) je měřena několika metodami: PART, JRip, J48graft a Bagging-PART. Nejlepším řešením se zdá být kombinace rysů 1 až 4 (RFQdiscRFQcompar, ARF, RDist, SDRD), která splňuje jak požadavek na nízký počet rysů, tak na vysokou úspěšnost.



Tabulka 5.4: Nejvhodnější kombinace atributů podle průměrné hodnoty evaluační míry *accuracy* ze čtyř metod: PART, JRip, J48graft a Bagging-PART. Nejlepším řešením se zdá být kombinace rysů 1 až 4 (RFQdiscRFQcompar, ARF, RDist, SDRD), která splňuje jak požadavek na nízký počet rysů, tak na vysokou úspěšnost.

atributy	průměrná hodnota accuracy
1	86.86
1-2	89.53
1-3	92.77
1-4	**93.42**
1-5	93.27
1-6	93.48
1-7	93.70
1-8	93.86
1-9	93.93
1-10	93.86
1-22	94.29

Tabulka 5.5: Feature ranking všech vlastností použitých při vyhledávání víceslovných termínů. Jednotlivé rysy jsou seřazeny podle důležitosti role, jakou hrají při automatickém vyhledávání víceslovných termínů. Feature ranking je vyhodnocen podle data-miningového nástroje Weka.

pořadí	zkratka rysu
1	MWT:t-score
2	MWT:T1
3	MWT:T2
4	MWT:MI-score
5	MWT:Oblig
6	MWT:Prox
7	MWT:AFC5
8	MWT:AFC3
9	MWT:Modus

Víceslovný termín je typem kolokace, konkrétně pravidelnou systémovou kolokací (Čermák, 2010). Lexikální asociační míry se používají při vyhledávání kolokací v textech, podobně jako některé kontextové vlastnosti (viz obligatornost a proximitu níže).

T-score je někdy označován jako míra kontrastu. Čím vyšší je hodnota t-score, tím méně je pravděpodobné, že jde o náhodné rozložení slov a tím je také pravděpodobnější, že jde o pevnou, ustálenou kombinaci slov. Zdálo by se tedy, že čím vyšší je hodnota t-score, tím vyšší je pravděpodobnost, že je daný bigram jakožto kolokace víceslovným termínem. V tomto případě je ale třeba vzít v úvahu negativní korelaci t-score a označení bigramu za víceslovný termín (viz kap. 5.2.2 a obrázek 5.4). Ta naznačuje, že pro vyhledání víceslovných termínů má význam spíš nižší hodnota t-score. Tato lexikální asociační míra totiž protežuje slova s velmi vysokou frekvencí (např. funkční slova), přičemž termíny obvykle mívají frekvenci v rámci korpusu spíše nižší. T-score tedy nepřispívá k označení bigramu za víceslovný termín jakožto kolokace, naopak odlišuje víceslovné termíny od kolokací jiného druhu (např. spojení frekventovaného slovesa s předložkou v rámci valence).

MI-score (angl. mutual information, vzájemná informace) bere v úvahu pravděpodobnost výskytu prvního slova, pravděpodobnost výskytu druhého a pravděpodobnost, že se obě slova vyskytnou najednou. MI-score na rozdíl od t-score klade větší důraz na slova s nízkou (až velmi nízkou) frekvencí. Čím vyšší je hodnota MI-score, tím vyšší je pravděpodobnost, že daný bigram je kolokací.

slovných pouze čtyři rysy); více v kap. 5.2.3.

Už samotný fakt, že zkoumaný bigram je kolokací, zvyšuje pravděpodobnost, že jde o víceslovný termín. Velké procento kolokací v akademických textech totiž tvoří právě víceslovné termíny.

Většina víceslovných termínů obsahuje alespoň jeden jednoslovný. Proto byly mezi rysy sloužící k vyhledání víceslovných termínů zařazeny vlastnosti **MWT:T1** a **MWT:T2**. Ty jsou jednotlivým bigramům přiřazovány automaticky na základě identifikace jednoslovných termínů v textech - pokud je první nebo druhá pozice bigramu automaticky identifikovaným termínem, je bigramu přidělen rys MWT:T1, resp. MWT:T2. Tento rys zvyšuje pravděpodobnost označení bigramu za víceslovný termín[15], zvláště pokud jde zároveň o ustálenou kolokaci.

Při vyhledávání víceslovných termínů jsou užitečné i kontextové vlastnosti vzájemná obligatornost kontextu a proximita, využívané mj. i pro vyhledávání kolokací v textech (p-kolokace; Cvrček, 2013). **MWT:Oblig** neboli vzájemná obligatornost kontextu zachycuje míru podmíněnosti vzájemného souvýskytu dvou členů bigramu. Sleduje počet výskytů členu B bigramu v blízkém kontextu členu A a naopak. Tento počet je vyjádřen jako procento z celkové frekvence členu A (resp. B). Čím vyšší je hodnota tohoto rysu, tím vyšší je i pravděpodobnost, že jde o víceslovný termín.

Proximita, **MWT:Prox**, je průměr absolutních hodnot textových vzdáleností dvou slov. Nabývá hodnot 0 (pro slova, která se nevyskytují v blízkém kontextu) nebo 1 a více (obv. 1 až 3). Čím vyšší je hodnota rysu MWT:Prox (v rozmezí 1-3), tím menší je pravděpodobnost, že jde o kolokaci, tedy o víceslovný termín. Pokud je hodnota rysu 0, pak nemůže jít o víceslovný termín, protože slova se nevyskytují v blízkém kontextu (k tomu ovšem v případě bigramů nemůže dojít).

Obecně můžeme na základě těchto poznatků o víceslovných termínech tvrdit, že pro jejich vyhledávání jsou nejdůležitější dva fakty: jsou to ustálené kolokace a jejich součástí bývá alespoň jeden jednoslovný termín.

[15] I bigramy obsahující jeden termín mohou být ale velmi silnými netermíny, záleží i na terminologické platnosti druhého členu bigramu a na hodnotách dalších atributů.

5.2.2 Korelace vlastností

Parametry experimentu 15: Korelace vlastností víceslovných termínů. V experimentu byla vytvořena korelační matice všech rysů používaných ve výzkumu, vč. ručně přiřazovaného rysu „MWT“ (víceslovný termín). Jako materiál jsou použita sloučená trénovací data (ALL). Materiál je v podobě bigramů, bez lemmatizace, s opakováním a bez slovních druhů. Hranice mezi termíny a netermíny je v nástroji Weka defaultně nastavena na hodnotu 0,5.

Jak se ukázalo už v případě lexikální asociační míry t-score (viz výše), k interpretaci výsledků feature rankingu je zapotřebí informace o tom, jakým způsobem spolu korelují jednotlivé atributy. Zjišťuje se síla korelace (slabá nebo silná korelace), a také kladná či záporná hodnota (kladná korelace: čím vyšší je hodnota A, tím vyšší je hodnota B; záporná korelace: čím vyšší je hodnota A, tím nižší je hodnota B). Vlastnosti, které spolu korelují, nemusejí jedna druhou způsobovat (Volín, 2007).

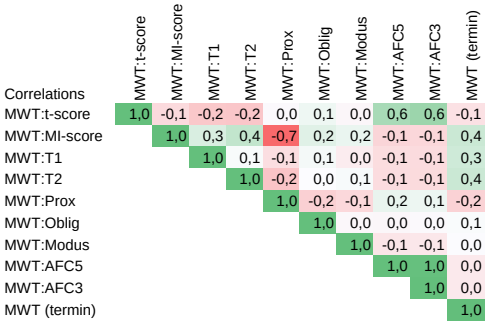
Znázornění korelací mezi jednotlivými atributy má podobu korelační matice (obrázek 5.4), z níž je zřejmé, zda korelace mezi dvěma konkrétními rysy je kladná (zelená barva), či záporná (červená barva), a jak vysokou má hodnotu (čím tmavší odstín, tím silnější korelace).

Pro zjištění, jakou roli hrají zkoumané rysy ve vyhledávání víceslovných termínů je důležitý především poslední sloupec korelační matice, který ukazuje, jaká je korelace jednotlivých rysů s označením bigramu za víceslovný termín. Kladná korelace naznačuje, že čím vyšší je hodnota dané vlastnosti, tím vyšší je pravděpodobnost, že se jedná o víceslovný termín. Záporná korelace ukazuje opačný vztah: čím vyšší je hodnota dané vlastnosti, tím nižší je pravděpodobnost, že daný bigram bude víceslovným termínem.

Čím je tedy např. hodnota MI-score vyšší, tím vyšší je pravděpodobnost, že jde o termín. Skutečnost, že druhé slovo v bigramu je termín, také zvyšuje pravděpodobnost víceslovného termínu. Naopak čím vyšší je hodnota t-score nebo proximity, tím méně je pravděpodobné, že jde o víceslovný termín.

Z ostatních korelací mezi samotnými rysy je zajímavá je například vysoká negativní korelace kontextové proximity s asociační mírou MI-score či vysoká pozitivní korelace kontextových rysů MWT:AFC5 a MWT:AFC3 s asociační mírou t-score.

Obrázek 5.4: Korelační matice použitých vlastností mezi sebou a s ručně označeným víceslovným termínem (poslední sloupec matice). Čím tmavší je barva, tím silnější je korelace mezi dvěma rysy, červená barva znamená kladou korelaci a zelená zápornou.



5.2.3 Nejvhodnější kombinace vlastností pro ATR

Parametry experimentu 16: Nejvhodnější kombinace vlastností víceslovných termínů. Experiment je zaměřen na nalezení vhodné kombinace atributů sloužící k automatickému vyhledávání termínů. Vhodnost kombinací se hodnotí na základě úspěšnosti metody Bagging-PART (z nástroje Weka) na sloučených trénovacích datech (ALL). Materiál je v podobě bigramů, bez lemmatizace, s opakováním a bez slovních druhů. Hranice mezi termíny a netermíny je v nástroji Weka defaultně nastavena na hodnotu 0,5.

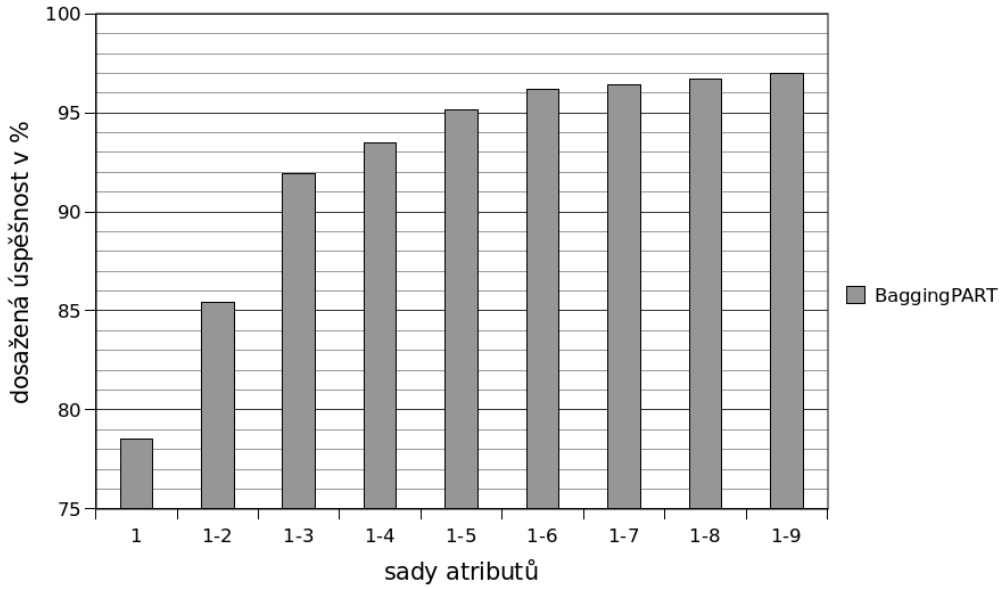
Stejně jako při automatickém vyhledávání termínů jednoslovných, i při vyhledávání víceslovných termínů je výhodné najít menší kombinaci atributů, která má ještě poměrně vysokou úspěšnost. Příprava i průběh experimentů je s takovou vhodnou kombinací vlastností daleko jednodušší. Vyhledání takové kombinace má navíc velký význam jako vodítko pro další metody automatického vyhledávání termínů.

Prvním krokem pro nalezení takové efektivní kombinace je seřazení atributů podle důležitosti (feature ranking, kap. 5.2.1), následuje vyhodnocení úspěšnosti na datech s různým počtem rysů.

V tomto experimentu je srovnávána úspěšnost metody Bagging-PART z data-miningového nástroje Weka při použití různých kombinací atributů podle tabulky 5.5. K prvnímu, nejdůležitějšímu atributu se postupně přidávají další rysy, takže vznikne 9 kombinací (1, 1-2, 1-3 až 1-9).

Obrázek 5.5 ukazuje, jak výrazně mění přidávání jednotlivých rysů úspěšnost metody Bagging-PART při vyhledávání víceslovných termínů. Až do přidání šestého rysu (MWT:Prox) úspěšnost narůstá minimálně o jedno procento, rozdíl mezi kombinací 1-6 a 1-7 nebo 1-9 není už tak zásadní. Proto je za nejvhodnější označena kombinace prvních šesti rysů (viz tab. 5.6).

Obrázek 5.5: Srovnání úspěšnosti vyhledávání termínů s různými kombinacemi atributů. Atributy jsou řazeny podle výsledků feature rankingu. Úspěšnost (hodnota míry *accuracy*) je měřena metodou Bagging-PART. Nejlepší se zdá být kombinace rysů 1 až 6, která splňuje požadavek na vysokou úspěšnost a relativně malý počet rysů.



Tabulka 5.6: Nejvhodnější kombinace atributů podle průměrné úspěšnosti (*accuracy*) metody Bagging-PART. Nejlepším řešením se zdá být kombinace rysů 1 až 6 (MWT:T1, MWT:T2, MWT:t-score, MWT:MI-score, MWT:Oblig a MWT:Prox), která splňuje jak požadavek na poměrně nízký počet rysů, tak na vysokou úspěšnost.

kombinace	úspěšnost	název přidaného rysu
1	78.55	MWT:t-score
1-2	85.42	MWT:T1
1-3	91.91	MWT:T2
1-4	93.51	MWT:MI-score
1-5	95.14	MWT:Oblig
1-6	**96.21**	**MWT:Prox**
1-7	96.44	MWT:AFC5
1-8	96.73	MWT:AFC3

5.3 Shrnutí

Za pomoci metod v data-miningovém nástroji Weka bylo jednotlivým zkoumaným rysům jednoslovných i víceslovných termínů přiřazeno pořadí, a to na základě toho, jak důležitá byla jejich role ve vyhledávání termínů.

Mezi nejdůležitější rysy jednoslovných termínů patří relativní frekvence v disciplíně oproti srovnávacímu korpusu (RFQdiscRFQcompar), relativní distribuce v akademických disciplínách (RDist), průměrná redukovaná četnost slova v korpusu (ARF) a směrodatná odchylka od relativní distribuce (SDRD). Pořadí jednotlivých vlastností hraje velkou roli ve výběru nejvhodnější kombinace pro automatické vyhledávání termínů. Taková kombinace by měla být účinná ve vyhledávání termínů, ale zároveň by neměla obsahovat příliš velké množství rysů. Z tohoto pohledu se jako velmi vhodná jeví kombinace rysů s rankem 1 až 4, tedy výše zmíněné rysy.

Vhodná kombinace vlastností víceslovných termínů je rozsáhlejší, jedná se o šest rysů označených za nejdůležitější pro proces vyhledávání termínů: lexikální asociační míry t-score (MWT:t-score) a MI-score (MWT:MI-score), vysoká automaticky přiřazená terminologická platnost první a/nebo druhé pozice bigramu (MWT:T1 a MWT:T2) a kontextové rysy vzájemná proximita kontextu (MWT:Prox) a obligatornost kontextu(MWT:Oblig).

Feature ranking sice dokáže atributy seřadit podle důležitosti pro proces automatického vyhledávání termínů, ale nedokáže určit, zda má daný rys pozitivní, nebo negativní vliv na hodnotu terminologické platnosti textové pozice (tedy na pravděpodobnost, že textová pozice je termín). Z tohoto pohledu je cenná korelace mezi jednotlivými rysy jednoslovných termínů, zvláště mezi automaticky označeným rysem „termín“ a všemi ostatními vlastnostmi. Pokud je korelace kladná, stoupá pravděpodobnost, že textová pozice je jednoslovný termín, spolu se stoupající hodnotou konkrétního atributu; pokud je záporná, je tento vztah opačný. Vysoká korelace, ať už kladná či záporná, ozřejmuje vztahy mezi některými dalšími rysy (např. mezi frekvencí a kontextovými atributy).

U víceslovných termínů se korelace mezi jednotlivými rysy a automaticky označenými víceslovnými termíny osvědčily zejména v případě t-score. Negativní korelace totiž naznačuje, že čím vyšší je míra t-score, tím nižší je pravděpodobnost, že bigram je víceslovným termínem. T-score je tedy schopen odlišit typ kolokací, které nebývají termíny (kolokace vysoce frekventovaných, např. funkčních slov), od víceslovných termínů, jejichž jednotlivé složky bývají méně frekventované.

Výsledky automatického označování jednoslovných i víceslovných termínů a netermínů ve velkém množství textů přinášejí množství nových poznatků o termínech, případně ověření předpokladů předchozích výzkumů[16] a jejich zasazení do kontextu. Nejdůležitějším výsledkem z hlediska cíle dizertační práce je především seřazení vlastností termínů podle závažnosti jejich role při automatickém vyhledávání termínů a vytvoření vhodné, tj. malé a efektivní, sady těchto vlastností.

[16]Není náhodou, že význam frekvenčních a distribučních rysů často zmiňují autoři zabývající se kvantitativními vlastnostmi termínů (kap. 1.2.2.1); výsledky experimentů tyto předpoklady plně potvrzují a řadí různé rysy odvíjející se od frekvence a distribuce jednotlivých slov podle jejich závažnosti.

Závěr

Hlavním cílem předkládané dizertační práce bylo získat nové poznatky o termínech, zvláště o jejich vlastnostech a chování v textech různých oborů zahrnutých do korpusu SYN2010. Relevantními kvantitativními vlastnostmi je možné doplnit současné popisy termínu jako základní terminologické jednotky.

Kvantitativním vlastnostem termínů jednoslovných i víceslovných je věnována kapitola 5. Ta se zaměřuje na to, které ze zkoumaných vlastností hrají největší roli v automatickém vyhledávání termínů v textech.

Pomocí data-miningového nástroje Weka byly vlastnosti uspořádány podle důležitosti role, kterou hrály při vyhledávání termínů. Z takto seřazených kvantitativních rysů byla vytvořena skupina vlastností podle dvou kritérií: 1. co nejmenší počet rysů a 2. zachování vysoké úspěšnosti. Zároveň tvoří základ účinné metody automatického vyhledávání termínů, a jejich nalezení má proto velký význam pro budoucí výzkumy v této oblasti.

Nejdůležitějšími vlastnostmi jednoslovných termínů jsou poměr relativní frekvence v dané disciplíně (v subkorpusu textů oboru) ku relativní frekvenci v neakademických textech (v subkorpusu COMPAR), průměrná redukovaná četnost (pravidelnost rozmístění v korpusu), relativní distribuce textové pozice v akademických disciplínách a směrodatná odchylka od relativní vzdálenosti jednotlivých výskytů slova.

Mezi nejvýznamnější rysy víceslovných termínů patří přítomnost jednoslovného termínu na první nebo druhé pozici zkoumané dvojice slov a vlastnosti vztahující se ke kolokačním vlastnostem této dvojice: asociační míry MI-score a t-score a kontextové vlastnosti – vzájemná obligatornost kontextu a proximita.

Na základě pečlivě zvolené kombinace vlastností je možné doplnit současné popisy termínu tak, aby odrážely kvantitativní vlastnosti termínů v reálných textech:

Jednoslovný termín (ať už samostatně stojící, nebo součást víceslovného) je slovo, které se v odborných textech daného oboru vyskytuje výrazně častěji než v textech neakademických, najdeme ho jen v malém počtu akademických disciplín, v celém korpusu (SYN2010) je nerovnoměrně rozložené a málo frekventované a rozestupy mezi jeho jednotlivými výskyty jsou nepravidelné.

Víceslovný termín je ustálená kolokace složená z méně frekventovaných slov, která obvykle obsahuje alespoň jeden termín jednoslovný.

Hodnoty nejdůležitějších vlastností termínů do určité míry odpovídají i intuitivně vnímané škále v terminologii. Čím je například termín častější v určitém oboru oproti neakademickým textům nebo v čím menším počtu oborů se vyskytuje, tím větší je pravděpodobnost, že mu intuitivně (stejně jako automaticky) bude připisována vysoká terminologická platnost. Využití principu škály vede mimo jiné k otázce, kam umístit hranici mezi automaticky vyhledanými termíny a netermíny. Odpověď je závislá na povaze a prioritách daného výzkumu, umístění hranice na prostřední hodnotě (0,5 na škále 0 až 1) se ukázalo jako vhodné pro metodu TERMIT a pro cíle představovaného výzkumu.

Podmínkou pro spolehlivé zjištění zásadních vlastností termínů i dalších poznatků o termínech je nalezení úspěšné metody automatického vyhledávání termínů (ATR). Analýza v kapitole 3 ukázala, že metoda TERMIT, která je založena na data miningu (vytěžování informací z velkých objemů dat), se při automatické identifikaci termínů osvědčila. Lze tedy předpokládat, že na jejím základě můžeme vyvozovat obecnější závěry o termínech, například o jejich vlastnostech.

Následující tabulka shrnuje údaje o úspěšnosti metody při vyhledávání jednoslovných i víceslovných termínů prostřednictvím evaluačních měr používaných k vyhodnocení výsledků data miningu[1].

	jednoslovné termíny	**víceslovné termíny (bigramy)**
accuracy	95 %	97 %
precision	85 %	81 %
recall	72 %	75 %

Metoda TERMIT je nejen úspěšná při vyhledávání termínů, ale navíc oproti dalším metodám ATR unikátní v několika ohledech:

- vyhledává a zkoumá termíny v téměř čtyřiceti velmi odlišných akademických disciplínách
- je schopna identifikovat jak jednoslovné, tak víceslovné termíny; při hledání víceslovných termínů využívá automaticky nalezené termíny jednoslovné

[1] Jednotlivé hodnoty v tabulce jsou založeny na automatickém vyhledávání termínů v trénovacích (hodnota míry *accuracy*) a v testovacích datech (*precision* a *recall*).

- je založena na data miningu, dokáže tedy využívat velmi složitých vztahů mezi zkoumanými vlastnostmi
- zkoumá velký počet vlastností známých z předchozích výzkumů a řadí je podle důležitosti; tím poskytuje informace cenné pro další výzkum v oblasti ATR.

Při zkoumání automaticky vyhledaných jednoslovných a víceslovných termínů v textech, kterému se věnuje kapitola 4, je třeba přistupovat opatrně k zobecňování poznatků. Každá akademická disciplína má totiž svá specifika a může se od ostatních do velké míry odlišovat. Poznatky o termínech jsou přitom ve velké míře založeny na datech ze čtyř oborů, na něž především byla zaměřena pozornost. Jde o obory informatika, literární věda, lékařství a sociologie (byly vybrány tak, aby se od sebe co možná nejvíce lišily).

Průměrný počet automaticky vyhledaných jednoslovných termínů v textech 37 akademických disciplín obsažených v korpusu SYN2010 je 22 % ze všech textových pozic (především slov a interpunkce). Jednotlivé obory se ale mezi sebou velmi liší – hodnoty se pohybují mezi 10 a 50 % textu obsazeného jednoslovnými termíny, ať už samostatně stojícími či součástmi víceslovných termínů. Menší počet termínů je obvyklý u oborů humanitních, kdežto disciplíny technické či přírodovědné mívají v textech vyšší množství termínů. Ze všech jednoslovných termínů stojí zhruba 60 % v textech samostatně, zbylých 40 % jednoslovných termínů jsou součástí víceslovných.

Víceslovné termíny obsazují v textech čtyř vybraných oborů zhruba 13 % textových pozic (obv. slov); z nich je zhruba 75 % tvořeno termíny jednoslovnými (nesamostatnými). Samostatných jednoslovných termínů se ve zkoumaných akademických textech vyskytuje více než termínů víceslovných, ve slovní zásobě jednotlivých oborů ale s vysokou pravděpodobností víceslovné termíny počtem mnohonásobně převyšují termíny jednoslovné.

Na základě automaticky vyhledaných termínů lze sledovat i další charakteristiky termínů, například slovní druhy jednoslovných termínů a složení termínů víceslovných, případně i vztahy mezi jednotlivými obory na základě sdílených termínů.

Nejdůležitějším výzkumným úkolem pro blízkou budoucnost je aplikace metody TERMIT na další jazyky, především angličtinu. Dále by bylo vhodné zaměřit pozornost i na skupinu oborově nespecifických vědeckých termínů, které by mohly být vyhledány podobným způsobem jako jednoslovné i víceslovné oborové termíny. Seznam takových termínů by mohl sloužit jako pomůcka při studiu na střední nebo vysoké škole, při výuce češtiny jako cizího jazyka či při sestavování slovníku češtiny.

Předkládaný výzkum ukazuje, že i čistě kvantitativní přístup, jako je data mining, je vhodný pro zkoumání lingvistického (korpusového) materiálu. S jeho pomocí lze s velkou úspěšností automaticky vyhledávat jevy (termíny) ve velkých objemech korpusových dat, a zároveň doplňovat teoretické popisy daných jevů.

Literatura

ISO 1087:1990. *Terminology – vocabulary.* Geneva: International Organization for Standardization, 1990.

ISO 1087-1:2000. *Terminology – vocabulary – part 1: Theory and application.* Geneva: International Organization for Standardization, 2000.

BEČKA, J. V. 1972. The lexical composition of specialized texts and its quantitative aspect. *Prague Studies in Mathematical Linguistics.* 1972, vol. 4, s. 47–64.

BOZDĚCHOVÁ, I. 2009. *Současná terminologie (se zaměřením na kolokační termíny z lékařství).* Praha: Karolinum.

CABRÉ, M. T. 1999. *Terminology: Theory, Methods and Applications.* Amsterdam/Philadelphia: John Benjamins Publishig Company.

CABRÉ, M. T. 2003. Theories of terminology: Their description, prescription and explanation. *Terminology.* 2003, vol. 9, no. 2, s. 163–199.

COXHEAD, A. 2000. A new academic word list. *TESOL Quarterly.* 2000, vol. 34, no. 2, s. 213–238.

CVRČEK, V. 2013. *Kvantitativní analýza kontextu.* Praha: Nakladatelství Lidové noviny/Ústav Českého národníko korpusu.

CVRČEK, V. 2014. Kvantitativní určení lexikálního jádra jazyka. *Časopis pro moderní filologii.* 2014, vol. 96, no. 1, s. 9–26.

CVRČEK, V.; KOVÁŘÍKOVÁ, D. 2011. Možnosti a meze korpusové lingvistiky. *Naše řeč.* 2011, vol. 94, no. 3, s. 113–133.

ČERMÁK, F. 2000. Jazykový korpus: Prostředek a zdroj poznání. In ČERMÁK, F.; KLÍMOVÁ, J.; PETKEVIČ, V. (ed.) *Studie z korpusové lingvistiky.* Praha: Karolinum, 2000, s. 15–37.

ČERMÁK, F. 2001. Termín a frazém: Případ překrývání a periférie dvou nominativních oblastí. In ŽEMLIČKA, M. (ed.). *Termina 2000.* Praha: Galén, 2001, s. 31–36.

ČERMÁK, F. 2010. *Lexikon a sémantika.* Praha: Nakladatelství Lidové noviny.

ČERMÁK, F.; KŘEN, M. (ed.). 2004. *Frekvenční slovník češtiny.* Praha: Nakladatelství Lidové noviny.

ČERMÁK, F.; KŘEN, M. 2011. *A Frequency Dictionary of Czech.* London: Routledge.

ČERMÁKOVÁ, A. 2009. *Valence českých substantiv.* Praha: Nakladatelství Lidové noviny.

Český národní korpus – SYN2010. 2010. Praha: Ústav Českého národního korpusu FF UK. Dostupný z WWW: <http://www.korpus.cz>.

Český národní korpus – SYN. 2014. Praha: Ústav Českého národního korpusu FF UK. Dostupný z WWW: <http://www.korpus.cz>.

DA SYLVA, L. 2009. Corpus-based derivation of a „basic scientific vocabulary“ for indexing purposes. In MAHLBERG, M.; GONZÁLEZ-DÍAZ, V. (ed.) *Proceedings of the CL Conference*. Dostupný z WWW: <http://ucrel.lancs.ac.uk/ publications/cl2009/>.

DUCHOŇ, J. 2001. Několik poznámek k pravopisu odborných výrazů v terminologii a nomenklatuře přírodních věd a medicíny. In ŽEMLIČKA, M. (ed.). *Termina 2000*. Praha: Galén, 2001.

FILIPEC, J.; ČERMÁK, F. 1985. *Česká lexikologie*. Praha: Academia.

FILIPEC, J. et al. 2005. *Slovník spisovné češtiny pro školu a veřejnost*. 4. vyd. Praha: Academia.

PALMER, F. R. 1968. *Selected Papers of J. R. Firth 1952-1959*. London: Longman.

FRANTZI, K.; ANANIADOU, S. 1996. Extracting nested collocations. In *Proceedings of the Sixteenth Conference on Computational Linguistics*. Copenhagen, 1996, s. 41–46.

FRANTZI, K.; ANANIADOU, S. 1997. Automatic term recognition using contextual cues. In *Proceedings of the 2nd Workshop on Multilinguality in Software Industry: the AI Contribution (MULSAIC'97)*. Nagoya, 1997, s. 73–80.

FRANTZI, K.; ANANIADOU, S. 1999. The C/NC value domain independent method for multi-word term extraction. *Journal of Natural Language Processing*. 1996, vol. 3, no. 2, s. 115–127.

GRIES, S. T. 2009. *Statistics for Linguistics with R*. Berlin/New York: Walter de Gruyter.

HALLIDAY, M. A. K.; TEUBERT, W.; YALLOP, C.; ČERMÁKOVÁ, A. 2004. *Lexicology and Corpus Linguistics. An Introduction*. London/New York: Continuum.

HARTMANN, R. R. K.; JAMES, G. 1998. *Dictionary of Lexicography*. London: Routledge.

HAUSENBLAS, K. 1962. K specifickým rysům odborné terminologie. In BĚLIČ, J.; DOLEŽEL, L.; PECIAR, Š. (ed.) *Problémy marxistické jazykovědy*. Praha: ČSAV, 1962, s. 248–262.

HYLAND, K.; TSE, P. 2007. Is there an „academic vocabulary“? *TESOL Quarterly*. 2007, vol. 41, no. 2, s. 235–253.

CHUNG, T. M. 2003. A corpus comparison approach for terminology extraction. *Terminology*. 2003, vol. 9, no. 2, s. 221–246.

JELÍNEK, J.; BEČKA, J. V.; TĚŠITELOVÁ, M. 1961. *Frekvence slov, slovních druhů a tvarů v českém jazyce.* Praha: SPN.

KAGEURA, K. 1997. A preliminary investigation of the nature of frequency distributions

of constituent elements of terms in terminology. *Terminology*. 1997, vol. 4, no. 2, s. 199—223.

KAGEURA, K.; DAILLE, B.; NAKAGAWA, H.; CHIEN, L. 2004. Recent trends in computational terminology. *Terminology*. 2004, vol. 10, no. 1, s. 1–21.

KAGEURA, K.; UMINO, B. 1996. Methods of automatic term recognition: A review. *Terminology*. 1996, vol. 3, no. 2, s. 259–289.

KAHOVEC, J. 2001. Chemická nomenklatura a lingvisté. In ŽEMLIČKA, M. (ed.). *Termina 2000*. Praha: Galén, 2001.

KARLÍK, P.; NEKULA, M.; PLESKALOVÁ, J. (ed.) 2002. *Encyklopedický slovník češtiny*. Praha: Nakladatelství Lidové noviny.

KIT, C.; LIU, X. 2008. Measuring mono-word termhood by rank difference via corpus comparison. *Terminology*. 2008, vol. 14, no. 2, s. 204–229.

KOCOUREK, R. 1965. Termín a jeho definice. *Československý terminologický časopis*. 1965, vol. 4, no. 1, s. 1–25.

KOPECKIJ, L. 1935. O lexikálním plánu hospodářského jazyka. *Slovo a slovesnost*. 1935, vol. 1, s. 120–122.

L'HOMME, M. 2006. The processing of terms in dictionaries: New models and techniques (a state of art). *Terminology*. 2006, vol. 12, no. 2, s. 181–188.

L'HOMME, M.; HEID, U.; SAGER, J. C. 2003. Terminology during the past decade (1994-2004): An editorial statement. *Terminology*. 2003, vol. 9, no. 2, s. 151–161.

LAURISTON, A. 1995. Criteria for measuring term recognition. In *EACL '95 Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics*. San Francisco: Morgan Kaufmann Publishers, 1995, s. 17–22.

LEMAY, C; L'HOMME, M.; DROUIN, P. 2005. Two methods for extracting „specific" single-word terms from specialised corpora: experimentation and evaluation. *International Journal of Corpus Linguistics*. 2005, vol. 10, no. 2, s. 227-55.

LINHART, J.; PETRUSEK, M.; VODÁKOVÁ, A.; MAŘÍKOVÁ, H. 1996. *Velký sociologický slovník*. Praha: Karolinum.

HALL, M. et al. 2009. The Weka data mining software: An update. *SIGKDD Explorations*. 2009, vol. 11, no. 1.

MACHAČ, J. 1964. Odborná terminologie ve výkladovém slovníku. *Československý terminologický časopis*. 1964, vol. 3, s. 65–76.

MACHOVÁ, S. 1995. Terminografie. In ČERMÁK, F; BLATNÁ, R. (ed.) *Manuál lexikografie*. Praha: H&H, 1995, s. 137–157.

MANNING, C.D.; SCHÜTZE, H. 2000. *Foundations of Statistical Natural Language Processing*. Cambridge/London: The MIT Press.

MARÍN, M.J. 2014. Evaluation of five single-word term recognition methods on legal English corpus. *Corpora*. 2014, vol. 9, no. 1, s. 83–107.

MCKILLUP, S. 2012. *Statistics Explained*. Cambridge: Cambridge University Press.

MÜLLER, R.; ŠIDÁK, P. 2012. *Slovník novější literární teorie*. Praha: Academia.

NENADIĆ, G.; ANANIADOU, S.; MCNAUGHT, J. 2004. Ehancing automatic term recognition through recognition of variation. In *Proceedings of COLING 2004*. Geneva, 2004, s. 604–610.

NOBATA, C.; COLLIER, N. H.; TSUJII, J. 1999. Automatic term identification and classification in biology texts. In *Proceedings of Natural Language Pacific Rim Symposium (NLPRS'99)*. 1999, s. 369–374.

NÜNNING, A.; TRÁVNÍČEK, J.; HOLÝ, J. 2006. *Lexikon teorie literatury a kultury*. Brno: Host.

OBDRŽÁLEK, J. 2001. Méně známé aspekty fyzikální terminologie. In ŽEMLIČKA, M. (ed.). *Termina 2000*. Praha: Galén, 2001, s. 31–36.

PECINA, P. 2009. *Lexical Association Measures: Collocation Extraction*. Praha: Institute of Formal and Applied Linguistics.

POŠTOLKOVÁ, B.; ROUDNÝ, M.; TEJNOR, A. 1983. *O české terminologii*. Praha: Academia.

SAVICKÝ, P.; HLAVÁČOVÁ, J. 2003. Measures of word commonness. *Journal of Quantitative Linguistics*. 2003, vol. 9, no. 3, s. 215–231.

SINCLAIR, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

SINCLAIR, J. 2004. *Trust the Text: Language, corpus and discourse*. London: Routledge.

ŠRAJEROVÁ, D. 2009a. Automatic term recognition as a resource for theory of terminology. In MAHLBERG, M.; GONZÁLEZ-DÍAZ, V. (ed.) *Proceedings of the Corpus Linguistics Conference*. Dostupný z WWW: <http:/ ucrel.lancs.ac.uk/publications/cl2009/>.

ŠRAJEROVÁ, D. 2009b. Automatické vyhledávání termínů a jeho dopad na definici termínu. *Časopis pro moderní filologii*. 2009, vol. 91, no. 1, s. 11–19.

ŠRAJEROVÁ, D.; KOVÁŘÍK, O.; CVRČEK, V. 2009. Automatic term recognition based on data-mining techniques. In *Proceedings of Computer Science and Information Engineering - CSIE*. Los Angeles, s. 453–457.

TEICH, E.; FANKHAUSER, P. 2010. Exploring a Corpus of Scientific Texts Using Data Mining. In *Corpus-linguistic Applications: Current Studies, New Directions*. Amsterdam:

Rodopi, 2010, s. 233–247.

TEUBERT, W. 2005a. Language as an Economic Factor: The Importance of Terminology In BARNBROOK, G.; DANIELSSON, P.; MAHLBERG, M. (ed.) *Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora.* London/New York: Continuum, 2005, s. 96–106.

TEUBERT, W. 2005b. My version of corpus linguistics. *International Journal of Corpus Linguistics.* 2005, vol. 10, no. 1, s.1–13.

TEUBERT, W.; ČERMÁKOVÁ, A. 2004. Directions in Corpus Linguistics. In HALLIDAY, M.A.K. (ed.) *Lexicology and Corpus Linguistics: An Introduction.* London/New York: Continuum, 2004, s. 113–165.

TOGNINI-BONELLI, E. 2001. *Corpus Linguistics at Work.* Amsterdam/Philadelphia: John Benjamins Publishig Company.

VILLE-OMETZ, F.; ROYAUTÉ, J.; ZASADZINSKI, A. 2007. Enhancing in automatic recognition and extraction of term variants with linguistic features. *Terminology.* 2007, vol. 13, no. 1, s. 35–59.

VIVALDI, J.; CABRERA-DIEGO, L.A.; SIERRA, G.; POZZI, M. 2012. Using Wikipedia to Validate the Terminology Found in Corpus of Basic Textbooks. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 12).* Dostupný z WWW: <http://www.lrec-conf.org/proceedings/lrec2012/ index.html>.

VLAŠÍN, V. (ed.) 1984. *Slovník literární teorie.* Praha: Československý spisovatel.

VOKURKA, M.; HUGO, J. et al. 2009. *Velký lékařský slovník.* Praha: Maxdorf.

VOLÍN, J. 2007. *Statistické metody ve fonetickém výzkumu.* Praha: Epocha.

WERMTER, J.; HAHN, U. 2005. Finding new terminology in very large corpora. In *Proceedings of the 3rd International Conference on Knowledge Capture (KCAP 2005)..* 2005, s. 137–144.

WEST, M. 1953. *A General Service List of English Words.* London: Longman, Green and Co.

WITTEN, I. H.; FRANK, E. 2005. 2. vyd. *Data Mining: Practical Machine Learning Tools and Techniques.* Amsterdam: Elsevier.

YANG, H. 1986. A new technique for identifying scientific/technical terms and describing science texts: An interim report. *Literary and Linguistic Computing.* 1986, vol. 1, no. 2, s. 93–103.

ZIPF, G. K. 1935. *The Psychobiology of Language.* Boston: Houghton Mifflin.

A Vysvětlivky

Accuracy Evaluační míra hodnotící podíl správných výsledků v populaci, ať už jde o instance příznakové, nebo nepříznakové; zde ukazuje, kolik termínů je správně označeno jako termíny a kolik netermínů správně je označeno jako netermíny. Accuracy je v rámci předkládaného výzkumu často použitá jako podklad pro hodnocení úspěšnosti konkrétní metody. Viz kap. 2.4.

Akademický subkorpus (korpus SCI) Subkorpus obsahující všechny texty z 37 oborů v korpusu SYN2010. Viz kap. 2.2.1.

ALL Zkratka pro sloučená trénovací data ze čtyř oborů (COM, LIT, MED, SOC). Celkem osm tisíc textových pozic. Viz kap. 2.2.3.

ATR (Automatic Term Recognition) Automatické vyhledávání termínů. Viz kap. 1.1.2.

Atribut Statistická nebo lingvistická vlastnost zkoumaná v rámci daného výzkumu, přiřazovaná jednotlivým textovým pozicím, resp. bigramům. Viz kap. 2.2.4.

AWL (Academic Word List) Seznam termínů bez tématické příslušnosti k danému oboru. Viz kap. 1.3.1.2.

Bagging-PART Metoda z data-miningového nástroje Weka. Viz kap. 2.3.4.1.

BayesNet Metoda z data-miningového nástroje Weka. Viz kap. 2.3.4.1.

Bigram Sřetězení dvou slov v textu.

COMPAR viz Korpus srovnávací.

Corpus-driven výzkum Výzkum řízený korpusem (lingvistickými daty). Viz kap. 1.1.3.

Data mining Proces hledání opakujících se vzorů ve velkých objemech dat za pomoci počítačových algoritmů. Viz kap. 2.1.

Feature ranking Seřazení atributů použitých v data miningu podle důležitosti. Viz kap. 5.1.1.

Feature selection Výběr několika nejdůležitějších atributů. Viz kap. 5.1.1.

Instance V data miningu jeden případ ze skupiny případů, které jsou zkoumány. Instanci zde odpovídá textová pozice v korpusu (buď ve formě slovního tvaru, nebo lemmatu). Viz kap. 2.1.

J48graft Metoda z data-miningového nástroje Weka. Viz kap. 2.3.4.1.

JRip Metoda z data-miningového nástroje Weka. Viz kap. 2.3.4.1.

Korelace Jeden z nástrojů statistického popisu: Dvě proměnné spolu korelují, pokud se jejich hodnoty mění společně v určitých kvantifikovatelných krocích. Viz kap. 5.1.2.

Korpus srovnávací (COMPAR) Korpus vytvořený k porovnání frekvencí textových pozic v akademickém subkorpusu a v textech neakademických. COMPAR je složený z textů beletristických a publicistických. Viz tab. 2.2 v kap. 2.2.2.

Lemma Reprezentativní slovníková podoba hesla.

Lemmatizace Automatický proces, při kterém je každé formě v korpusu přiděleno lemma.

Matice Rozsáhlá tabulka, kde v řádcích jsou jednotlivé instance (textové pozice) a ve sloupcích jsou instancím přiřazovány hodnoty jednotlivých atributů (vlastností). Viz kap. 2.1.

Míra lexikální asociační Matematický postup (vzorec) používaný pro vyhledání kolokací v korpusu. Viz kap. 2.2.4.

Míra statistická evaluační Matematický postup (vzorec) používaný pro vyhodnocení výsledků experimentů. Mezi zde používané míry patří accuracy, precision, recall a F-measure. Viz kap. 2.4.

Nejasné viz tabulku seznam atributů (atribut N) nebo kap. 2.2.4

Obligatornost Vzájemná obligatornost kontextu. Atribut používaný k vyhledávání víceslovných termínů. Podmíněnost vzájemného souvýskytu dvou členů bigramu. Viz kap. 2.2.4.

PART Metoda z data-miningového nástroje Weka. Viz kap. 2.3.4.1.

Polyfunkčnost Mezioborová polysémie. Viz kap. 1.3.1.5.

Precision Evaluační míra hodnotící podíl relevantních instancí ze všech vyhledaných instancí; zde ukazuje, kolik procent textových pozic označených jako termíny jsou skutečně termíny.

Proximita Atribut používaný k vyhledávání víceslovných termínů. Průměr absolutních hodnot textových vzdáleností dvou slov. Viz kap. 2.2.4.

Recall Evaluační míra hodnotící podíl vyhledaných instancí, které jsou relevantní; zde ukazuje, kolik procent ze všech termínů bylo skutečně vyhledáno.

Rys viz Atribut

Subkorpus Podmnožina zdrojového korpusu. Viz kap. 2.2.2.

Termín jednoslovný Pokud není rozlišeno, pak jednoslovný termín samostatně stojící i součást víceslovného termínu.

Termín jednoslovný samostatný/samostatně stojící Jednoslovný termín, který stojí v textu samostatně - není tedy součástí termínu víceslovného.

Termín oborově nespecifický Zvláštní případ termínu z oblasti filozofie - teorie vědy, statistiky apod., který se stal součástí obecnější akademické slovní zásoby. V rámci akademických textů má vysokou frekvenci i distribuci, ale obvykle je bez tématické příslušnosti k danému oboru. Viz kap. 1.3.1.2.

Termín víceslovný Termín složený z několika slov, obv. aspoň z jednoho termínu jednoslovného.

Terminologická platnost Míra, do jaké je dané slovo termínem nebo netermínem. Hodnota terminologické platnosti (na škále 0 až 1) se přiřazuje každé z textových pozic. Hranice mezi termíny a netermíny je zde stanovana na hodnotu 0,5 (za termíny jsou považovány textové pozice s terminologickou platností 0,5 a více).

Testovací data Vstupní data, na nichž se testuje, jak úspěšně je nástroj schopen vyhledávat termíny v nových datech. Zde všechny textové pozice ze 37 akademických oborů (textový typ označený SCI) v korpusu SYN2010. Všechny pozice (celkem 9 milionů pozic) byly ve výzkumu automaticky zařazeny mezi jednoslovné termíny, nebo mezi netermíny.

TERMIT (Term Mining Tool) Metoda automatického vyhledávání termínů založená na data miningu. Viz kap. 2.3.1.

Textová pozice V korpusové terminologii grafické slovo oddělené z obou stran mezerou či interpunkční značky (které při sazbě často bezprostředně sousedí se slovem). Termín textová pozice se používá v souvislosti s tím, že každý text, který vstupuje do korpusu, prochází procesem tokenizace.

Trénovací data Vstupní data, na nichž data-miningový nástroj vytvoří pravidla pro stanovení terminologické platnosti a natrénuje se na ně. Zde vybraná a podrobně zpracovaná data ze 4 oborů (COM, LIT, MED, SOC). Z každé disciplíny bylo vybráno 2000 po sobě jdoucích textových pozic, každá z nich byla ručně označena jako termín či netermín, případně jako součást víceslovného termínu. Sloučená data, tj. osm tisíc textových pozic ze čtyř různých oborů, mají zkratku ALL.

Vstup/vstupní data Data ve formě matice (tabulky) poskytnutá data-miningovému nástroji pro učení. Viz kap. 2.1.

Výstup/výstupní data Výsledky procesu data miningu: nalezená pravidla, pomocí nichž je možné přiřadit hodnotu terminologické platnost k libovolné textové pozici; výstupem může být i ohodnocení důležitosti jednotlivých rysů pro proces učení nebo ohodnocení úspěšnosti použité metody. Viz kap. 2.1.

Weka Data-miningový nástroj, který nabízí práci s velkým množstvím metod, z nichž některé dokážou řadit použité atributy podle důležitosti (viz feature ranking) nebo vyhledat vhodnou kombinaci atributů (viz feature selection). Více v kap. 2.3.3.

ZeroR Nejjednodušší klasifikační metoda, která označí všechny instance jako příslušníky majoritní třídy. Zde označí všechny textové pozice jako netermíny. Používá se k porovnávání výsledků složitějších metod. Viz kap. 2.3.4.1 a 2.4.

Značkování korpusu morfologické Proces, při kterém jsou textovým pozicím v korpusu automaticky přidělovány morfologické značky. Ty jsou sumarizací gramatické informace o hledaném slovu (pozici) v konkrétním kontextu.

B Seznam zkoumaných vlastností

ARF	průměrná redukovaná četnost
CaseL	lemma začíná malým písmenem
CaseU	lemma začíná velkým písmenem (vlastní jména)
KontextE	kontext: entropie
KontextHgen	vážený průměr relativních frekvencí předcházejícího kontextu
KontextHH	kontext: Herfindahl-Hirschmanův index
Len_{syl}	délka slova ve slabikách
MWT:AFC3	absolutní frekvence kolokátu v kontextu -3 až 3 (pro vícesl.)
MWT:AFC5	absolutní frekvence kolokátu v kontextu -5 až 5 (pro vícesl.)
MWT:MI-score	asociační míra MI-score (pro víceslovné termíny)
MWT:Modus	modus výskytu (pro víceslovné termíny)
MWT:Oblig	vzájemná obligatornost kontextu (pro víceslovné termíny)
MWT:Prox	proximita (pro víceslovné termíny)
MWT:t-score	asociační míra t-score (pro víceslovné termíny)
MWT:T1	první pozice je jednoslovný termín (pro víceslovné termíny)
MWT:T2	druhá pozice je jednoslovný termín (pro víceslovné termíny)
NoCompar	instance se vůbec nevyskytuje v COMPAR
RARF	relativní průměrná redukovaná četnost
RDist	relativní distribuce v disciplínách
RFQ_{compar}	relativní frekvence v COMPAR
RFQ_{disc}	relativní frekvence v disciplíně
$\text{RFQ}_{disc}\text{RFQ}_{compar}$	poměr relativních frekvencí v disciplíně a v COMPAR
RFQ_{sci}	relativní frekvence ve SCI
$\text{RFQ}_{sci}\text{RFQ}_{compar}$	poměr relativních frekvencí ve SCI a v COMPAR
SDRD	směrodatná odchylka relativní vzdálenosti sousedních výskytů
SDRFQ	standardní odchylka relativních frekvencí v disciplínách
Struct	ne/obvyklost struktury slova

C Zkratky oborů

AGR	zemědělství, lesnictví
ANT	antropologie
ART	výtvarné umění
BIO	obecná biologie
BOT	botanika
CHE	chemie
CIN	film
COM	informatika
ECO	ekonomie, obchod
EDU	pedagogika
ENE	energetika
ENV	ekologie
ETH	etnografie
GEO	geologie
HIS	historie
IND	průmysl, technika
INF	knihovnictví, informace
JUR	právo
LIN	lingvistika
LIT	literární věda
LOG	logika
MAN	management, řízení
MAT	matematika
MED	lékařství
MIL	vojenství
MUS	hudba
NAT	jiný z oblasti přírodních věd
PHI	filozofie
PHY	fyzika
POL	politologie
PSY	psychologie
REL	náboženství, teologie
SOC	sociologie
SPO	sport
THE	divadlo, balet
TRA	doprava, telekomunikace
ZOO	zoologie

D Automaticky vyhledané termíny

Texty ze všech oborů zahrnutých do výzkumu s automaticky přidělenými hodnotami terminologické platnosti jsou k dispozici na přiloženém DVD.

E Vzorce

RFQ$_{disc}$ (Relativní frekvence v disciplíně)

Hodnota relativní frekvence se vypočítá jako poměr frekvence výskytů instance v textech daného oboru ku počtu všech textových pozic v oboru (jak je uveden v tab. 2.1). To je vyjádřeno vzorcem

$$RFQ_{disc}(x) = \frac{FQ_{disc}(x)}{N_{disc}},$$

kde $FQ_{disc}(x)$ je frekvence x instance v příslušné disciplíně a N$_{disc}$ je počet textových pozic v disciplíně.

RFQ$_{sci}$ (Relativní frekvence v subkorpusu SCI)

Hodnota relativní frekvence se vypočítá jako poměr frekvence výskytů instance v textech subkorpusu akademických textů (SCI) ku počtu všech textových pozic v subkorpusu SCI (celkový počet pozic ve všech oborech v tab.2.1). Vzorec pro výpočet RFQ$_{sci}$ je následující:

$$RFQ_{sci}(x) = \frac{FQ_{sci}(x)}{N_{sci}},$$

kde $FQ_{sci}(x)$ je frekvence instance x v subkorpusu SCI a N_{sci} je celkový počet textových pozic v subkorpusu SCI.

RFQ$_{compar}$ (Relativní frekvence v subkorpusu COMPAR)

Hodnota relativní frekvence se vypočítá jako poměr frekvence výskytů instance v textech subkorpusu neakademických textů COMPAR ku počtu všech textových pozic v subkorpusu (jak je uveden v tab. 2.2). Může se stát, že se daná instance v subkorpusu COMPAR nevyskytuje vůbec, potom je jí přidělena hodnota 0.

Pokud je hodnota frekvence instance ve srovnávacím korpusu COMPAR nenulová, spočítáme hodnotu RFQ$_{compar}$ podle vzorce

$$RFQ_{compar}(x) = \frac{FQ_{compar}(x)}{N_{compar}},$$

kde $FQ_{compar}(x)$ je frekvence instance v subkorpusu COMPAR a N_{compar} je počet textových pozic v COMPAR.

RFQ$_{disc}$RFQ$_{compar}$ (Poměr relativních frekvencí disciplína/obecné texty)

Hodnota tohoto atributu se vypočítá jako poměr relativní frekvence instance v dané disciplíně ku relativní frekvenci téže instance v korpusu neakademických textů COMPAR. Zjednodušeně řečeno jde o to, jak často se instance vyskytuje v daném oboru oproti korpusu obecnému. Pokud se daná instance vůbec nevyskytuje v obecném korpusu COMPAR, atribut RFQ$_{disc}$RFQ$_{compar}$ má hodnotu nekonečno. To zkresluje celkové výsledky, proto se z technických důvodů těmto instacím přiřazuje speciální hodnota, která je jasně odliší od ostatních instancí, ale výsledky přitom nezkreslí. Pro tyto případy je vytvořen doplňkový atribut NoCompar[1].

Čím vyšší je hodnota tohoto atributu, tím větší je nepoměr mezi frekvencí v textech daného oboru a frekvencí v textech obecných (v neprospěch obecných textů).

Hodnota RFQ$_{disc}$RFQ$_{compar}$ se počítá pomocí vzorce

$$RFQ_{disc}RFQ_{compar}(x) = \frac{RFQ_{disc}(x)}{RFQ_{compar}(x)}$$

(k RFQ_{disc} a RFQ_{compar} viz výše).

RFQ$_{sci}$RFQ$_{compar}$ (Poměr relativních frekvencí odborné texty/obecné texty)

Hodnota tohoto atributu se vypočítá jako poměr relativní frekvence instance v subkorpusu akademických textů SCI ku relativní frekvenci téže instance v obecném korpusu COMPAR. Je to tedy informace o tom, jak často se instance vyskytuje v odborném korpusu oproti korpusu obecnému. Pokud se daná instance vůbec nevyskytuje v obecném korpusu COMPAR, atribut RFQ$_{sci}$RFQ$_{compar}$ má hodnotu nekonečno. To zkresluje celkové výsledky, proto se z technických důvodů těmto instacím přiřazuje speciální hodnota, která je jasně odliší od ostatních instancí, ale výsledky přitom nezkreslí. Pro tyto případy je vytvořen doplňkový atribut NoCompar[2].

Čím vyšší je hodnota tohoto atributu, tím větší je nepoměr mezi frekvencí v odborných textech a frekvencí v textech obecných (v neprospěch obecných textů).

[1] NoCompar má hodnotu 1, vyskytuje-li se daná instance pouze v odborných textech, a nikoli v korpusu obecných textů COMPAR (pokud se instance vyskytne i v COMPAR, hodnota atributu NoCompar je 0.

[2] NoCompar má hodnotu 1, vyskytuje-li se daná instance pouze v odborných textech, a nikoli v korpusu obecných textů COMPAR (pokud se instance vyskytne i v COMPAR, hodnota atributu NoCompar je 0.

Hodnotu $\mathrm{RFQ}_{sci}\mathrm{RFQ}_{compar}$ vypočítáme podle vzorce

$$RFQ_{sci}RFQ_{compar}(x) = \frac{RFQ_{sci}(x)}{RFQ_{compar}(x)}$$

(k RFQ_{sci} a RFQ_{compar} viz výše).

NoCompar (Nulový výskyt v subkorpusu COMPAR)

Doplňkový atribut k poměru relativních frekvencí. Pokud se daná instance vyskytuje v subkorpusu odborných textů, ale má nulovou frekvenci v obecném korpusu COMPAR, je hodnota atributu NoCompar 1, ve všech ostatních případech je hodnota 0. Jde o zjednodušení atributu RFQ_{compar}, které je důležité pro proces učení data-miningového nástroje.

RDist (Relativní distribuce)

Základem pro výpočet hodnoty tohoto atributu je počet oborů v subkorpusu SCI, v kterých se daná instance vyskytuje. Tento počet je vydělen počtem všech oborů v subkorpusu SCI, atribut tedy nabírá hodnoty v intervalu 0-1.

Hodnotu RDist vypočítáme podle vzorce

$$RDist(x) = \frac{Dist(x)}{N_{disc}},$$

kde Dist je počet oborů, kde se vyskytuje daná instance a N_{disc} je celkový počet oborů v subkorpusu SCI.

SDRFQ (Směrodatná odchylka $\mathbf{RFQ}_{disc}$) Výsledek automatického zpracování korpusových dat. Směrodatná odchylka pro relativní frekvenci dané instance ve všech oborech. Jde o údaj vypovídající o distribuci instance v různých oborech, ovšem přesnější než RD, protože se bere v úvahu i frekvence výskytů.

Čím menší je hodnota směrodatné odchylky, tím je frekvence instance v oborech rovnoměrnější, zjednodušeně by se dalo říct, že distribuce dané instance je vyšší.

Hodnotu RDist vypočítáme podle vzorce

$$SDRFQ(x) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N_{disc}}(RFQ^{i}_{disc}(x) - AVG(RFQ_{disc}(x)))^2},$$

přičemž

$$RFQ^d_{disc}(x) = \frac{FQ^d_{sci}(x)}{FQ_{compar}(x)},$$

kde FQ^d_{sci} je frekvence slova v oboru d ze subkorpusu akademických textů (SCI) a $AVG()$ je průměr přes všechny disciplíny.

RARF (Relativní průměrná redukovaná četnost) Relativní průměrná redukovaná četnost je ARF vydělená frekvencí. Je výhodnější než samotná průměrná redukovaná četnost v tom ohledu, že potlačuje silný vliv frekvence. Pomocí ARF můžeme srovnávat pouze instance s podobnou frekvencí, díky RARF lze srovnávat i instance frekvenčně velmi odlišné.

RARF se vypočítá podle vzorce

$$RARF(x) = \frac{ARF(x)}{FQ(x)},$$

kde $ARF(x)$ je průměrná redukovaná četnost dané instance, jejíž hodnota je převzata z korpusu SYN2010 a $FQ(x)$ je frekvence instance v korpusu SYN2010.

SDRD (Směrodatná odchylka relativní vzdálenosti sousedních výskytů) Podobně jako ARF, jemuž se SDRD do určité míry podobá, se hodnoty vypočítávají pro celý korpus SYN2010. Jde o relativní odchylku vzdáleností mezi jednotlivými výskyty vydělených frekvencí dané instance. Smyslem SDRD je zjistit, jak pravidelně jsou jednotlivé výskyty rozmístěny v korpusu.

Čím vyšší je hodnota tohoto atributu, tím menší je rovnoměrnost rozmístění dané instance v korpusu (jednotlivé výskyty jsou ve shlucích).

SDRD se vypočítá podle vzorce

$$SDRD(x) = \sqrt{\frac{1}{N-1}\sum_{v \in V(x)} (d_v - \bar{d})^2},$$

kde V(x) je množina všech výskytů textové pozice x v korpusu a d_v je vzdálenost od výskytu v k dalšímu výskytu textové pozice.

KontextHH (Herfindahl-Hirschmanův index) Herfindahl-Hirschmanův index měří diverzitu jevů a zároveň i rovnoměrnost zastoupení jevů ve zkoumaném kontextu,

zohledňuje tedy počet typů i proporce. V rámci textu se může měřit levý i pravý kontext, ale vzhledem k tomu, že pravý kontext je obecně méně variabilní (poskytuje tedy o dané instanci více informací) (Cvrček 2012, s. 89), bude zde zohledňován pouze pravý kontext, konkrétně pozice +1.

Hodnota tohoto atributu se pohybuje v intervalu 0-1, přičemž čím vyšší je hodnota atributu KontextHH, tím je jev uspořádanější.

Herfindahl-Hirschmanův index se v počítá podle vzorce

$$KontextHH = \sum_{i=1}^{N} s_i^2,$$

kde p_i je pravděpodobnost *i-tého* kontextového typu vyjádřitelná jako podíl jeho absolutní frekvence na dané kontextové pozici a celkové frekvence klíčového slova.

Pro účely srovnávání slov s různým počtem kontextových typů (hodnota ACV) je třeba zvolit variantu normalizovanou, která nabývá hodnot od 0 do 1 a není tedy závislá na tomto ukazateli:

$$KontextHH^* = \frac{KontextHH - 1/N}{1 - 1/N}.$$

KontextE (Entropie) Výsledek automatického zpracování korpusových dat. Entropie je míra neuspořádanosti, v našem případě míra neuspořádanosti kontextu. Ze stejného důvodu jako u atributu KontextHHI, totiž že pravý kontext je méně variabilní (Cvrček 2012, s. 89), se zkoumá pouze pozice +1.

Čím nižší je hodnota tohoto atributu, tím je kontext uspořádanější.

Entropie se počítá podle vzorce

$$E = -\sum_{cfq \in CFQ} \frac{cfq}{FQ} log_2 \frac{cfq}{FQ},$$

kde CFQ je množina frekvencí všech slov v kontextu -1 a 1 od zkoumaného slova a FQ je frekvence zkoumaného slova.

Hgen (Vážený průměr relativních frekvencí předcházejícího kontextu) Atribut Hgen představuje vážený průměr relativních frekvencí slov v kontextu bezprostředně předcházejících zkoumanému slovu. Počítá se se dvěma až pěti předcházejícími slovy a podle toho se atribut značí Hgen2 až Hgen5.

Čím méně frekventovaná slova se v kontextu předcházejícím zkoumanému slovu vyskytují, tím je hodnota atributu vyšší.

Vzorec pro výpočet tohoto atributu je

$$Hgen^p(x_0) = AVG(-\sum_{i=-p}^{-1} RFQ_{SYN2010}(x_i) log_2 RFQ_{SYN2010}(x_i),$$

kde x_i je slovo na pozici i vzhledem ke zkoumanému slovu a AVG je průměr přes všechny výskyty daného slova.

MWT:MI (MI-score pro vyhledávání víceslovných termínů) Výsledek automatického zpracování korpusových dat. Jde o atribut používaný pouze při vyhledávání víceslovných termínů. MI-score (Mutual information, vzájemná informace) měří vzájemnou závislost dvou náhodných proměnných, v lingvistice jde o stanovení pravděpodobnosti, že se jedno slovo vyskytne v kontextu druhého (www.korpus.cz). V korpusové lingvistice je MI-score obvykle využíváno k vyhledávání kolokací.

MI-score se měří pro všechny bigramy v trénovacích datech. Počítá se podle vzorce

$$MI_{(xy)} = log_2 \frac{N f_{(xy)}}{f_{(x)} f_{(y)}}.$$

MWT:t-score (t-score pro vyhledávání víceslovných termínů) Výsledek automatického zpracování korpusových dat. Jde o atribut používaný pouze při vyhledávání víceslovných termínů. Vychází ze statistické metody testování hypotéz pomocí tzv. t-testu. Testuje se, jak hodně výskyty dvojic slov (po sobě jdoucích instancí) odpovídají náhodnému rozložení slov v korpusu (www.korpus.cz).

T-score se měří pro všechny bigramy v trénovacích datech. Počítá se podle vzorce

$$T(xy) = \frac{f_{(xy)} - \frac{f_{(x)} f_{(y)}}{N}}{\sqrt{f_{(xy)}}}.$$